TRANSPORTER : Transferring Visual Semantics from VLM Manifolds

Supplementary Material

Algorithm 1 ρ-OT

Input:
Embeddings $\widehat{\mathbf{z}}_{\Omega_1}, \widehat{\mathbf{z}}_{\Omega_2}$
Projection vectors $\{\mathbf{p}_{\Omega_1,\rho}\}_{\rho=1}^{P}, \{\mathbf{p}_{\Omega_2,\rho}\}_{\rho=1}^{P}$
Temp. τ, Reg. strength λ, Sinkhorn iter. K
Uniform $|\mathbf{N}| \times |\mathbf{N}|$ matrix $\mathbf{U}$ (where $\mathbf{U}_{i,j} = 1/|\mathbf{N}|$)
Output:
Transported $\tilde{\mathbf{z}}_{\Omega} = \tilde{\mathbf{T}}\widehat{\mathbf{z}}_{\Omega_1}$

$\mathbf{T} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathbf{N}|\times|\mathbf{N}|}$
for $\rho \in \{1, \dots, P\}$ **do**
  $\mathbf{a}_\rho \leftarrow \langle \widehat{\mathbf{z}}_{\Omega_1,i}, \mathbf{p}_{\Omega_1,\rho} \rangle$
  $\mathbf{b}_\rho \leftarrow \langle \widehat{\mathbf{z}}_{\Omega_2,j}, \mathbf{p}_{\Omega_2,\rho} \rangle$
  $\mathbf{M}_\rho \leftarrow \langle -\|\mathbf{a}_{i,\rho} - \mathbf{b}_{j,\rho}\|_2/\tau \rangle$ ▷ Neg. dist
  $\mathbf{M}_{\text{max}} \leftarrow \langle \max_j(\mathbf{M}_{\rho,i,j}) \rangle$
  $\mathbf{R}_\rho \leftarrow \langle \exp(\mathbf{M}_{\rho,i,j} - \mathbf{M}_{\text{max},i}) \rangle$
  $\mathbf{T}_\rho \leftarrow \text{diag}(\mathbf{R}_\rho \mathbf{1})^{-1}\mathbf{R}_\rho$ ▷ Row softmax
  $\mathbf{T}_{\rho,\lambda} \leftarrow (1-\lambda)\mathbf{T}_\rho + \lambda\mathbf{U}$ ▷ Uniform reg
  $\mathbf{T} \leftarrow \mathbf{T} + \mathbf{T}_{\rho,\lambda}$
end for
$\tilde{\mathbf{T}} \leftarrow \mathbf{T}/P$ ▷ Avg slice plans
for $k \in \{1, \dots, K\}$ **do** ▷ Sinkhorn-Knopp
  $\tilde{\mathbf{T}} \leftarrow \text{diag}(\tilde{\mathbf{T}}\mathbf{1})^{-1}\tilde{\mathbf{T}}$ ▷ Row norm
  $\tilde{\mathbf{T}} \leftarrow \tilde{\mathbf{T}}\text{diag}(\mathbf{1}^T\tilde{\mathbf{T}})^{-1}$ ▷ Column norm
end for
return $\tilde{\mathbf{T}}\widehat{\mathbf{z}}_{\Omega_1}$

6. Entropic Optimal Transport Formulation

Transport plan γ_ρ is found by minimizing transport cost $\mathbf{M} = |\mathbf{a} - \mathbf{b}|$ and maximizing entropy:

$$\min_{\gamma_\rho} \underbrace{\int_{\mathbb{R}\times\mathbb{R}} \mathbf{M} d\gamma_\rho(\mathbf{a},\mathbf{b})}_{\text{Transport cost}} - \tau \underbrace{\int_{\mathbb{R}\times\mathbb{R}} \gamma_\rho(\mathbf{a},\mathbf{b}) \log(\gamma_\rho(\mathbf{a},\mathbf{b})) d\mathbf{a} d\mathbf{b}}_{\text{Shannon entropy}} \quad (9)$$

where the differential $d\gamma_\rho(\mathbf{a},\mathbf{b})$ is the probability mass from $\mathbf{a} \to \mathbf{b}$ alogside entropic regularization. A discrete form of (9) for $\mathbf{N}$ tokens and transport plan $\mathbf{T}_\rho \in \mathbb{R}^{|\mathbf{N}|\times|\mathbf{N}|}$ over cost matrix $\mathbf{M}_{i,j,\rho} = \|\mathbf{a}_{i,\rho} - \mathbf{b}_{j,\rho}\|_2$, given temperature τ, can be formulated to find optimal transport $\tilde{\mathbf{T}}$:

$$\tilde{\mathbf{T}} = \underset{\mathbf{T}_\rho}{\text{argmin}} \frac{1}{P} \sum_{\rho=1}^{P} \Big(\underbrace{\sum_{i,j\in\mathbf{N}} \mathbf{T}_{i,j,\rho}\, \mathbf{M}_{i,j,\rho}}_{\text{Transport cost}} + \tau \underbrace{\sum_{i,j\in\mathbf{N}} \mathbf{T}_{i,j,\rho} \log(\mathbf{T}_{i,j,\rho})}_{\text{Entropy regularization}} \Big) \quad (10)$$

Solving the *full* doubly-constrained problem in (10) is computationally intensive. Thus, the proposed learnable ρ-OT provides an efficient approximation.

Table 4. **VLM architecture settings**. VideoLLaMA 3 (V), Gemma 3, (G), and Phi 4 MM (P), use different vision encoders, input resolutions, multi-modal projectors, and backbone LLMs.

VLM	Vision Encoder $\mathcal{E}$		MM proj	LLM	
	Model	res	dim	Model	prms
V	SoViT-400m/14 [2]	384^2	3584	Qwen2.5 [5]	7B
G	SigLIP-so400m/14 [92]	896^2	3584	Gemma 3 [87]	12B
P		$[28^2, 448^2]$	3072	Phi 4 Mini [1]	5B

7. Procedural details for OT

Projected embeddings $\widehat{\mathbf{z}}_{\Omega_1}$ are learned to minimize their mean divergence to target $\mathbf{z}_\Omega$. However, this does not guarantee that target local token structures are learned similar to those of $\widehat{\mathbf{z}}_{\Omega_2}$ optimized with the Gram-matrix loss. To combine their properties, learnable vectors $\mathbf{p}_{\Omega_1}, \mathbf{p}_{\Omega_2}$ are used to compute an optimal transport plan as shown in Algorithm 1. Embeddings $\widehat{\mathbf{z}}_{\Omega_1}, \widehat{\mathbf{z}}_{\Omega_2}$ are projected to $\mathbf{a}, \mathbf{b}$ based on their inner product with with $\mathbf{p}_{\Omega_1,\rho}, \mathbf{p}_{\Omega_2,\rho}$. Cost matrix $\mathbf{M}_\rho$ per ρ is computed from their negative pair-wise l2 distance, scaled by the temperature τ. This avoids computing a single, high-dimensional cost matrix and instead computes multiple, simpler cost matrices. The cost also enforces $\mathbf{p}_{\Omega_1}, \mathbf{p}_{\Omega_2}$ to produce projections with strong correspondence across the diagonal $\mathbf{a}_{i,\rho} \approx \mathbf{b}_{i,\rho}$. For numeric stability, the cost matrix $\mathbf{M}_\rho$ is normalized by row-wise softmax for P. To avoid sparsity, $\tilde{\mathbf{T}}$ is regularized with a uniform distribution $\frac{1}{|\mathbf{N}|}\mathbf{U}$ with weight λ. The final transport plan $\tilde{\mathbf{T}}$ is computed by averaging the sliced plans, followed by $K = 3$ Sinkhorn-Knopp [52] iterations to ensure a better approximation to the doubly-stochastic matrix of (10).

8. Additional implementation details

VLM selection. An overview of the selected VLM details is provided in Tab. 4. As shown, the models represent a wide spectrum of design choices varying in fundamental components such as visual encoders, embedding space size, and backbone language models. This heterogeneity aims to better reflect the general applicability of TRANSPORTER , as the presented results are not bound to a single model family or design paradigm, enabling a comprehensive analysis.

Coupling network datasets. In total, 200K videos are used for training the coupling network. This includes 40K VATEX, 80K LAVIB, and 80K Ego4D annotated clips. As VATEX is the smallest of the three, all training videos are used. In contrast, 160K total videos from LAVIB and Ego4D are selected. The choice of 200K iterations and dataset distri-

Table 5. **FVD and CLIPScores across dataset settings**. Evaluation metrics are reported as in Tab. 1. Best results are in **bold** and the used setting is in blue.

		EGO4D			
		40K	80K	120K	160K
LAVIB	40K	2.12e^2/34.86	1.72e^2/35.10	1.33e^2/35.41	1.23e^2/35.43
	80K	1.65e^2/35.24	1.25e^2/35.44	1.22e^2/35.50	1.20e^2/35.52
	120K	N/A	1.23e^2/35.50	1.19e^2/35.53	**1.18e^2/35.54**
	160K	N/A	1.19e^2/**35.54**	N/A	N/A

butions is found by the grid search in Tab. 5. Larger dataset samples only provide marginal improvements over the coupling network's reconstruction task, but with larger computational overheads. Thus, an 80/80K split is selected while taking into account training setup ablations.

Concept bank attributes. Tab. 6 presents learned vectors of concept/attribute modulations in context to the prompts used. In total, 33 unique modulations are presented in the paper, of which 10 relate to active objects, 11 are based on the action performed, 8 on scene changes, and 4 include object/action/scene combinations. With respect to the models used, 12 unique modulations are presented for Video LLaMA 3 logits, 11 for Gemma 3, and 10 for Phi 4 MM. The inclusion of different modulations shows the wide applicability of *TRANSPORTER* and its usability as a tool for exploring different aspects of video understanding.

9. OT types and compute times

TRANSPORTER is based on an optimal transport coupling between embedding spaces. Dimension size differences between representations limit the potential application of standard optimal transport approaches. As an alternative to the proposed ρ-OT module in Sec. 3.2, other learned approaches such as Wasserstein Auto-Encoders (WAE) [89], and Gromov-Wasserstein Auto-Encoders (GWAE) [66] can also be used. Tab. 7 presents FVD video quality across optimal transport methods as well as their computational overheads. Performance drops significantly for the heuristic approach as the transport plan in sGW does not preserve the sequential structure of tokens and their dynamics. The autoencoder [66, 89] approaches are implemented with two-MLP-layer encoders and decoders to avoid OOM errors. They achieve comparable performance to the proposed ρ-OT, but with a significant increase in computations required per iteration. Overall, ρ-OT with $P = 100$ provides a balanced OT approach given the compute overhead and performance.

10. Ablations with Phi 4 MM logits

Supplementary to Tab. 3, ablations are performed on *TRANSPORTER* settings on Phi 4 MM logits. Tab. 8 presents results over both coupling network and concept banks variants. Caption similarity drops in the inference-only setting compared to any of the other coupling network alternatives. Similarly, to Tab. 3, a small divergence is observed between the number of projections P. Overall, $P = 100$ still maintains competitive caption similarity with the average difference between the two top-performing projection settings being ± 0.71 across BLEU scores, ± 0.91 for CIDEr, ± 0.48 for METEOR, and ± 0.73 for SPICE. For the concept bank, most significant improvements are observed for CIDEr with $+2.31$ and B@2 with $+1.83$.

11. Modulations across phenomena

Critically, the applicability of VLMs depends on their ability to distinguish between opposing or distinct semantic meanings. Prior benchmarks have explored VLMs' response sensitivity across object plurality [24], preposition [23], and plausibility [22]. In parallel with these studies on VLM outputs, *TRANSPORTER* can provide visual representations of token predictions. Such modulation pairs linguistically ground VLMs across various aspects. This provides a novel path for future work to visually explore interpretability-related themes.

Plurality/counting. Fig. 9a presents modulation over `one` and `two` flowers being picked. VLMs often struggle with counting since it requires semantically matching language to visual prompts, both at a semantic level and in terms of location in frames. Videos present an additional challenge, as transitioning from `one` to `two` flowers requires an extra action that not all models can always infer.

Preposition. Fig. 9b visualizes a linguistic preposition, which should result in significant changes in the scene. The initial `woman shouting intensively to a man` is not successfully changed to `man shouting intensively to a woman`. The resulting videos from VideoLLaMA 3 and Phi 4 MM show that such fine-grained preposition details are often missed.

Relation/plausibility. Fig. 9c shows VLMs' reliance on priors in terms of object relations and physical plasibilities. Resulting videos when modulating to `sitting under a chair` either visualize the chair at a different location in the scene (lack of spatial understanding) or the scene resembles `sitting under a table` (potentially closer to the training distribution).

12. Additional qualitative results

In addition to the examples in Sec. 4.3, modulations for objects, actions, scenes, and their combinations are shown.

Active object. As in Figs. 11a and 11d, *TRANSPORTER* can generate videos to visualize transitions between different object attributes, such as color or material types. The sharpness of the transition also varies across modulations.

Table 6. **Prompts and modulations used for exploring active object, action performance, scene context, and combined settings logits**. Source-to-target modulations across Video LLaMA 3 (v), Gemma 3 (G), and Phi 4 MM (P) logits are denoted between π^- and target π^+.

VLM	Fig.	prompt (A ...)	Modulations π^-	Modulations π^+
		Active Objects		
G	3	close up shot of a [] bowling ball hitting pins in a bowling alley	red	blue
V	6	person jugging [] outdoors.	balls	clubs
V	11a	person pushing a [] stroller outdoorns during a sunny day.	purple	pink
V	11b	video of a scuba diver swimming underwater in the ocean and picking up a [].	starfish	seashell
V	11c	close-up of a person baking [] cookies. The cookies are on a baking tray.	christmas tree	circle
G	11d	video of a person writing TRANSPORTER on a [] table.	wood	metal
G	11e	video of a person reading a [] while sitting on a park bench.	book	newspaper
G	11f	person picking a(n) [] in a grocery store with shelves of fruits and vegetables in the background.	mango	apple
P	11g	person painting a wall with a [].	roll	brush
P	11h	top-view video of a person typing on a [].	typewriter	keyboard
		Action performance		
V	5	video of a person [] to get on the train.	walking	running
G	6	artistic gymnast performing a routine landing a dismount with a [] handspring and a full twist.	front	back
P	7c	close-up video of a person [] a car door with a cloth.	wiping	spraying
V	12a	video of a surfer [] at the open sea.	surfing	kite surfing
V	12b	a video of a teenager [] at a public staircase in a city center.	roller skating	skateboarding
V	12c	person [] in a lush field full of flowers	running	spinning
G	12d	video of a chef [] pizza dough in his kitchen	rolling	stretching
G	12e	video of a person that [] a cup of coffee in the kitchen.	stir	pour
G	12f	a cowboy [] a brown horse in the outback..	riding	leading
P	12g	person [] a soccer ball at a soccer field.	toe bouncing	head bouncing
P	12h	video of a person that [] boxes to his car.	picks up	drags
		Scene context		
V	2	video of a band playing music on a promenade. The camera zooms to the musician on the [] of the scene.	left	right
P	6	hitchhiker standing on the side of the road and puts up a cardboard sign that says [].	Chicago	New York
V	13a	person existing a classic car parked at a [].	parking lot	underpass
V	13b	golfer hitting the ball while on [].	grass	sandtrap
G	13c	person painting a wall [].	outdoors	indoors
G	13c	video of a tennis player getting ready to serve the ball on a [] court.	clay	grass
P	13e	group of firefighters rescuing a cat stranded on a [].	tree	roof
P	13f	video showing a group of people playing beach volleyball on a day with [].	overcast	clear sky
		Combined modulations		
P	7a,7b	chef cutting [] into [] pieces.	two, thin	one, thick
V	14a	person riding a [] while wearing a [] and [] jeans.	bicycle, jacket, black	scooter, t-shirt, blue
G	14b	video of a person with [] hair color during a [] walk around the city.	black, night	red, morning
P	14c	person [] at [] wearing an orange life vest.	canoeing, lake	kayaking, river

Table 7. **Video quality, semantic alignment, and compute times across OT methods**. Times are averaged over 1K videos across 3 runs on a single NVIDIA L40s with Video LLaMA 3 embeddings as targets. It is assumed that tensors are on the GPU without CPU offloading. Best results are in **bold** and the used setting is in blue.

OT Method		FVD↓	Compute time (secs. [secs./it])↓ fwd	bwd	tot
		Heuristic			
sGW $P=$	100	5.42e^2	272.48 [2.18]	N/A	272.48 [2.18]
		Learned			
WAE† $C=$	768	1.73e^2	538.81 [4.31]	178.67 [1.43]	717.43 [5.74]
	1152	1.46e^2	778.84 [6.23]	234.86 [1.88]	1012.03 [8.11]
GWAE† $C=$	1152	1.50e^2	853.73 [6.83]	217.55 [1.74]	1071.24 [8.57]
		TRANSPORTER			
ρ-OT $P=$	50	1.62e^2	**206.23** [**1.65**]	**49.97** [**0.40**]	**256.23** [**2.05**]
	100	1.25e^2	227.64 [1.82]	65.11 [0.52]	292.46 [2.34]
	200	1.09e^2	260.03 [2.08]	126.25 [1.01]	386.24 [3.09]
	400	**9.81e^1**	361.27 [2.89]	157.52 [1.26]	518.76 [4.15]

In fine-grained object changes such as [starfish] to [seashell] in Fig. 11b or [roll] to [brush] in Fig. 11g, the transition from π^- to π^+ happens over small $\Delta\omega$ fractions. Learned geometric and appearance object properties can also be seen during transitions in Figs. 11c and 11e with edges in complex shapes such as those of [christmas tree] being flattened to round edges, or the hardback of a [book] transitioning to a paperback, and eventually [newspaper] pages.

Action performance. In actions where their performance is proximal, such as [surfing] and [kitesurfing] in Fig. 12a, [roller skating] and [skateboarding] in Fig. 12b, and [toe bouncing] and [head bouncing] in Fig. 12g, modulations show that small and distinct changes are learned. The difference in speed can be seen in cases where action verbs differ significantly in their executions, such as [running] and [spinning] in Fig. 12c. Transitions between actions with even larger differences in their executions are simultaneously visible as in [riding] to [leading] in Fig. 12f.

Scene context. Scene modulations include backgrounds Figs. 13a, 13c, 13e and 13f, and locations Figs. 13b and 13d. The resulting videos show ***TRANSPORTER*** 's ability to generate videos with in-context modulations. Aspects such as the performance of the action or the appearances of objects/actors remain constant throughout the transition between source and target logits.

Combined. In settings with combined modulation, transitions of individual logit pairs are visualized over different $\Delta\omega$. Fig. 14a shows that transitioning between [black] and [blue] happens first, as their logit and embedding distance is shorter than other pairs. In contrast, the changes from [jacket] to [t-shirt], and [bicycle] to

† Trained on half batch size and double gradient accumulation steps due to OOM. Approx. fwd/bwd/tot reported.

Figure 10. **Modulation types comparisons** across VideoLLaMA 3, Gemma 3, Phi 4 MM, over (a) **plurarity/counting**, (b) **preposition**, and (c) **relations/plausibility**. The same rng (0) is used for *TRANSPORTER* inference across VLMs.

Table 8. **Phi 4 MM cosine similarity** (∇cos), **BLEU (B@1, B@2, B@3, B@4), CIDEr (C), METEOR (M), and SPICE (S)** over captions from *TRANSPORTER* videos and target captions. Settings are grouped in relation to the coupling network or the concept bank. Best results are **bold** and second best are underlined.

Method	Phi 4 MM							
	∇cos	B@1	B@2	B@3	B@4	C	M	S
Baseline [82]	0.04	25.22	11.10	4.35	0.01	5.01	11.84	8.64
Coupling network ablations								
inference-only	0.34	29.25	20.52	14.72	10.28	23.89	15.92	21.50
mean$\langle\Phi_{\Omega_1,\Omega_2}\rangle$	0.37	30.93	21.23	15.45	11.64	24.11	16.63	24.37
sGW OT	0.32	25.61	16.47	13.28	9.81	23.44	13.15	12.63
$P = \{$ 50	0.41	34.03	22.59	16.66	12.82	28.64	20.30	27.75
400	**0.44**	**35.24**	**23.78**	**18.53**	**14.08**	**30.42**	**21.35**	**28.96**
Concept bank ablation								
$\Delta\omega$ JS	0.36	33.15	21.42	15.87	11.93	28.20	18.56	27.42
TRANSPORTER	0.43	34.77	23.25	17.30	13.49	29.51	20.87	28.23

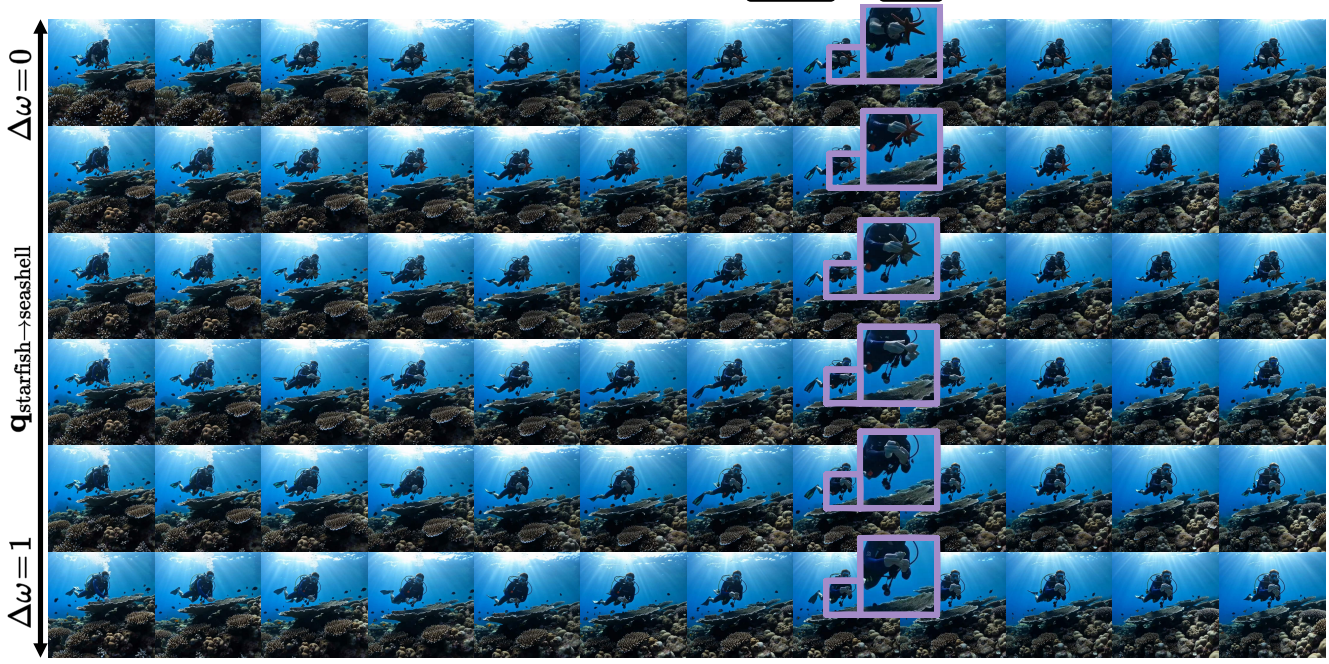
scooter happen at similar $\Delta\omega$ variance, showing that appearance attributes are encoded over larger manifolds.

13. Further discussions

Limitations. The manually selected modulations across model logits aim to visualize the learned representations and manifold between concepts. However, as noted, the interpolation is done based on learned vectors for concept pairs over VLM logit divergences. This can limit the clarity of explanations of entangled concepts for which vectors $\mathbf{q}_o$ are learned from source and target concept/attributes that are not directly orthogonal. As the current method is based on discrete concept pairs, a future direction could be generalizing L2V and *TRANSPORTER* to continuous attributes. For example, instead of slow $\leftrightarrow$ fast, the model could learn a speed vector, allowing a user to control the pace with a continuous scalar, not just an interpolation. This can further enable the applicability of *TRANSPORTER* in zero-shot concept pairs at inference.

Potential improvements. Although the focus has been on establishing L2V as a fruitful direction for video model interpretability, extending the method to few-shot video concept visualization can be an area of improvement. Instead of a discrete, static concept bank $\mathbf{Q}$, L2V can be formulated to model concept directions from a (learned) continuous function $f_\theta(\pi^-, \pi^+) \rightarrow \mathbf{q}_o$. This function could be trained to approximate the embedding manifold of any two text prompts used as input and output the corresponding latent direction vector $\mathbf{q}$. Potentially, this could be modeled contrastively [11, 92] on a large corpus of text-pair-video data. Inference-time semantic edits [18], uncertainty quantification [104], and model merging with model-specific vectors [35, 97] can also be adopted to reduce reliance on training concept vectors.

(a) Video LLaMA 3 modulations between `purple` and `pink`.



(b) Video LLaMA 3 modulations between `starfish` and `seashell`.

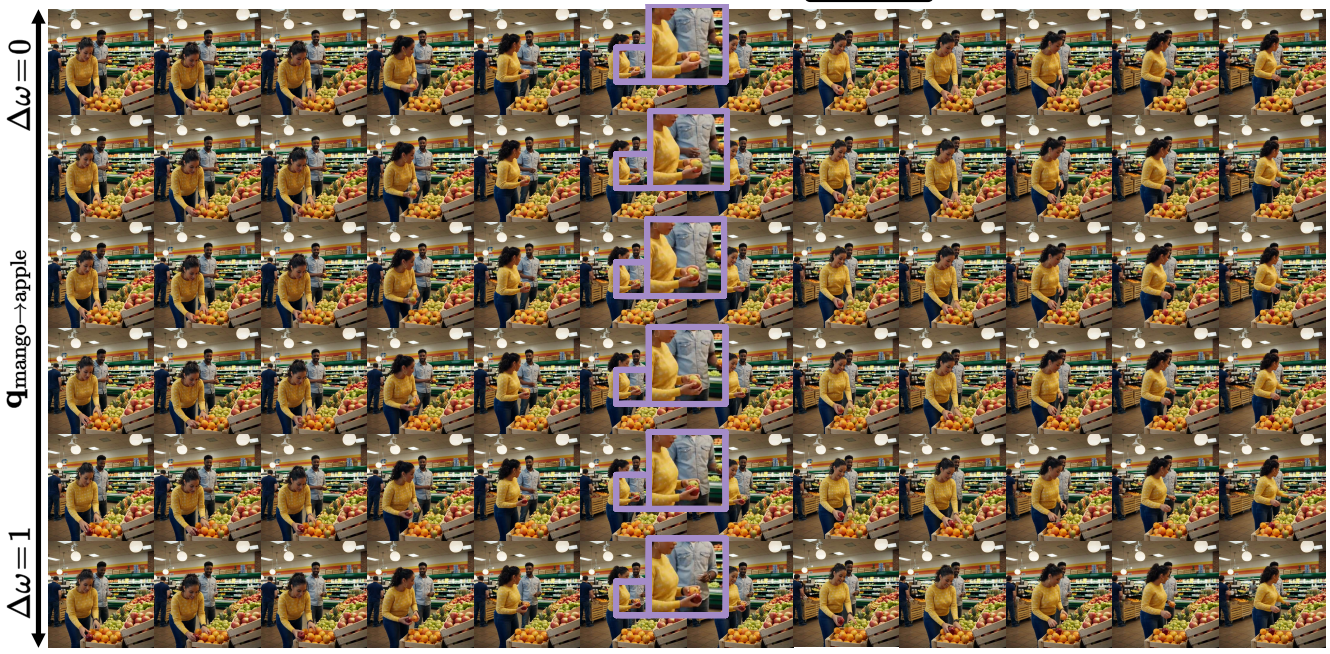
Figure 11. **Examples of active object modulations**. Frame quality is compressed due to filesize (best viewed digitally).

(c) Video LLaMA 3 modulations between `christmas tree` and `circle`.

(d) Gemma 3 modulations between `wood` and `metal`.

Figure 11. **Examples of active object modulations**. Frame quality is compressed due to filesize (best viewed digitally).

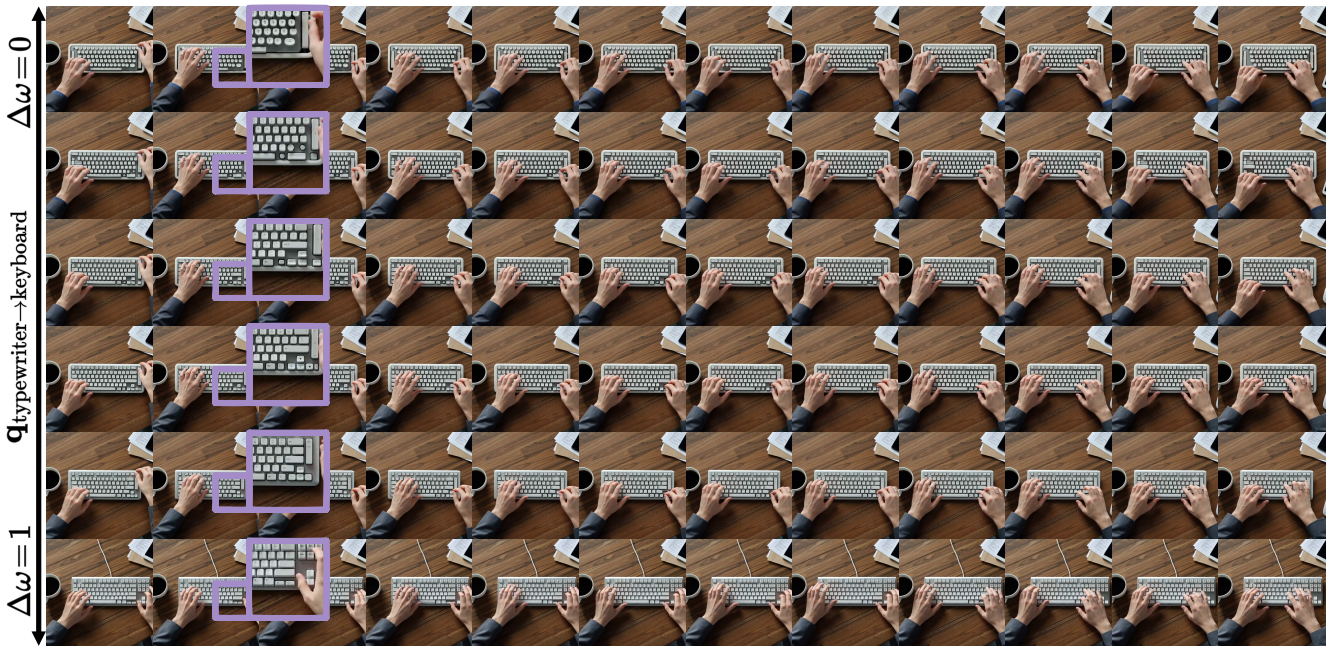
(e) Gemma 3 modulations between `book` and `newspaper`.



(f) Gemma 3 modulations between `mango` and `apple`.

Figure 11. **Examples of active object modulations**. Frame quality is compressed due to filesize (best viewed digitally).

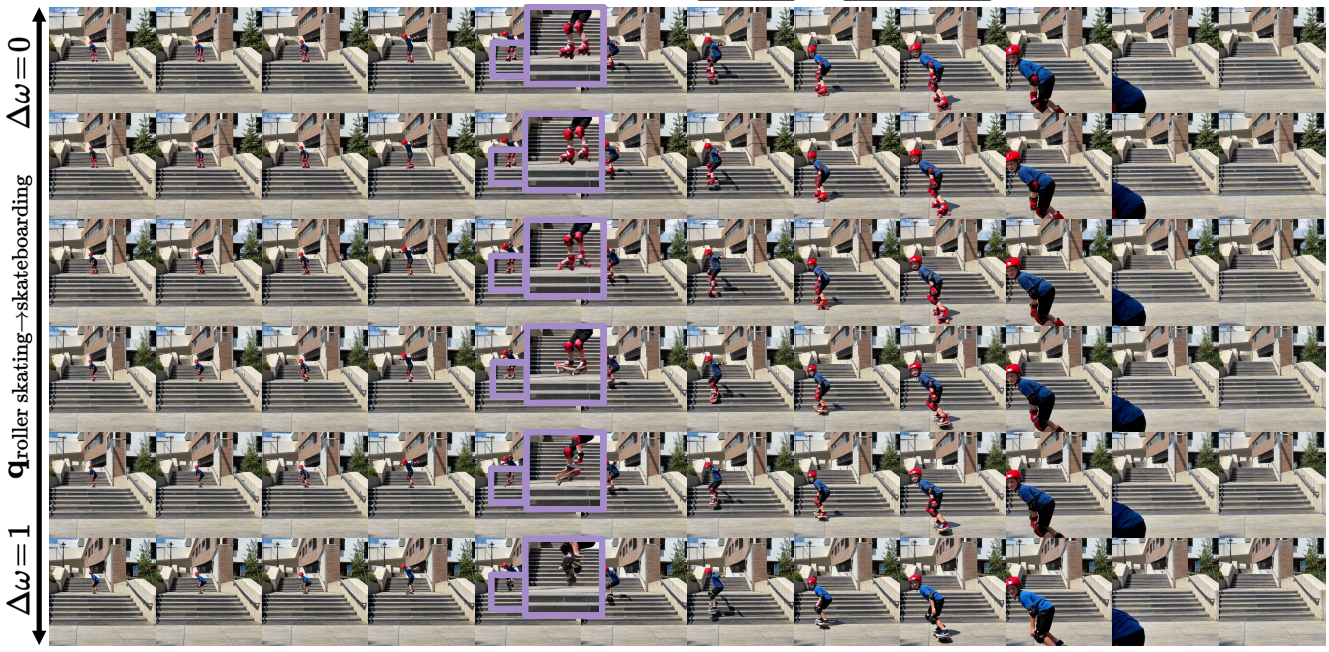
(g) Phi 4 MM modulations between `roll` and `brush`.



(h) Phi 4 MM modulations between `typewriter` and `keyboard`.

Figure 11. **Examples of active object modulations**. Frame quality is compressed due to filesize (best viewed digitally).

(a) Video LLaMA 3 modulations between `surfing` and `kitesurfing`.



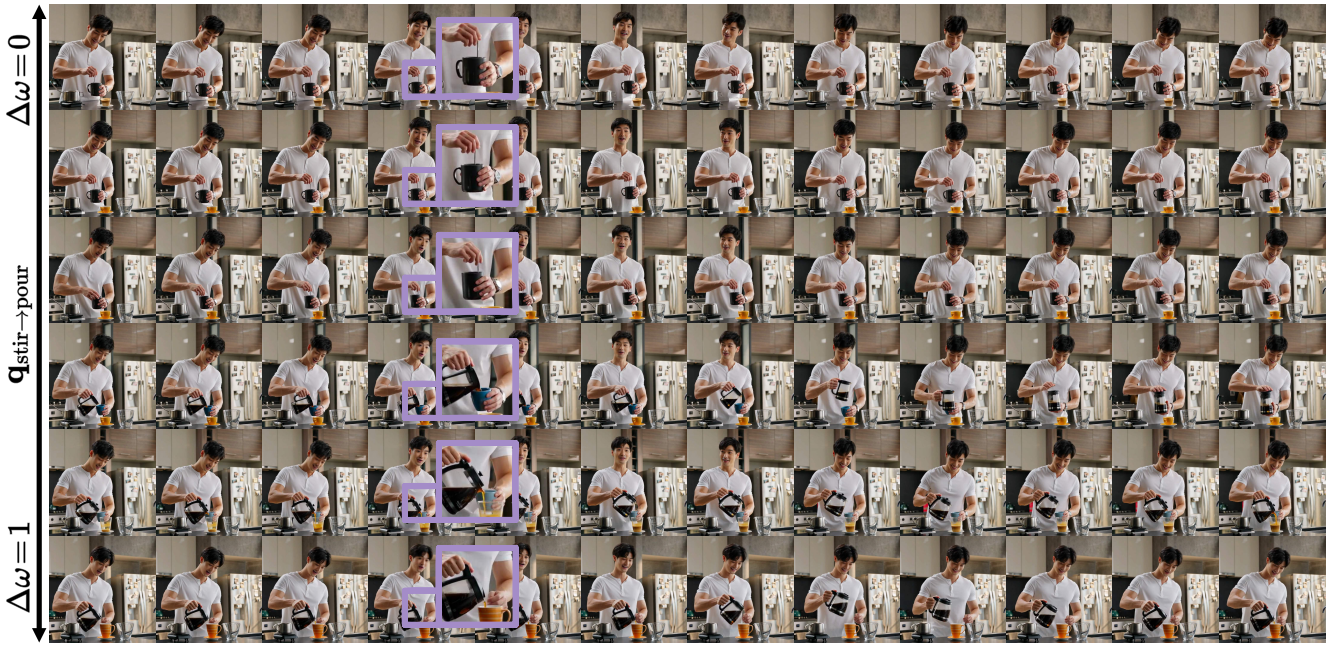
(b) Video LLaMA 3 modulations between `roller skating` and `skateboarding`.

Figure 12. **Examples of action modulations**. Frame quality is compressed due to filesize (best viewed digitally).

(c) Video LLaMA 3 modulations between `running` and `spinning`.

(d) Gemma 3 modulations between `rolling` and `stretching`.

Figure 12. **Examples of action modulations**. Frame quality is compressed due to filesize (best viewed digitally).

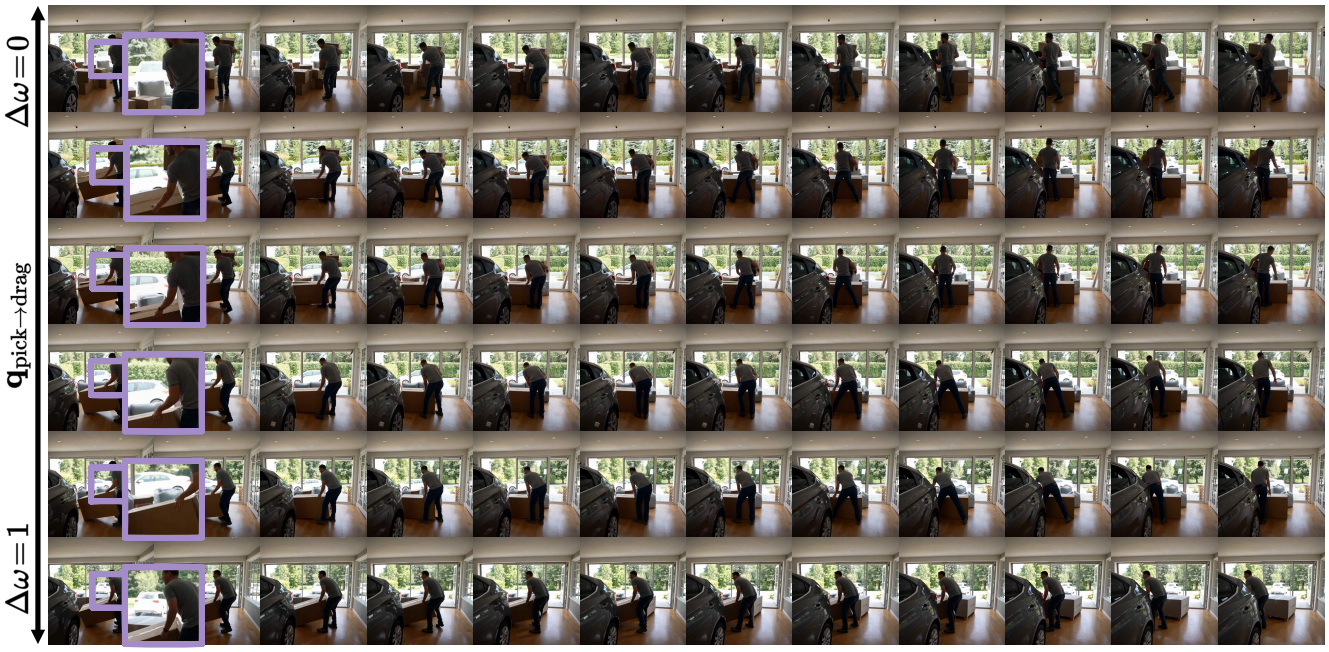


(e) Gemma 3 modulations between `stir` and `pour`.

(f) Gemma 3 modulations between `riding` and `leading`.

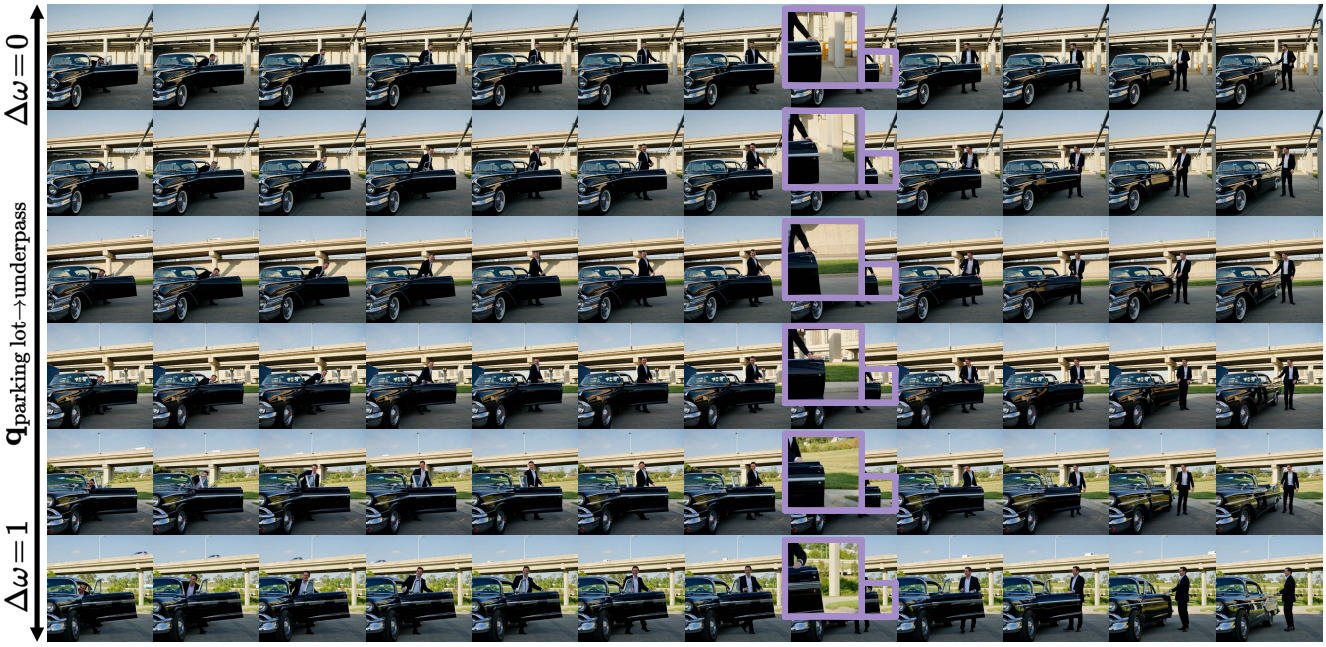
Figure 12. **Examples of action modulations**. Frame quality is compressed due to filesize (best viewed digitally).

(g) Phi 4 MM modulations between `toe bouncing` and `head bouncing`.



(h) Phi 4 MM modulations between `picks up` and `drags`.

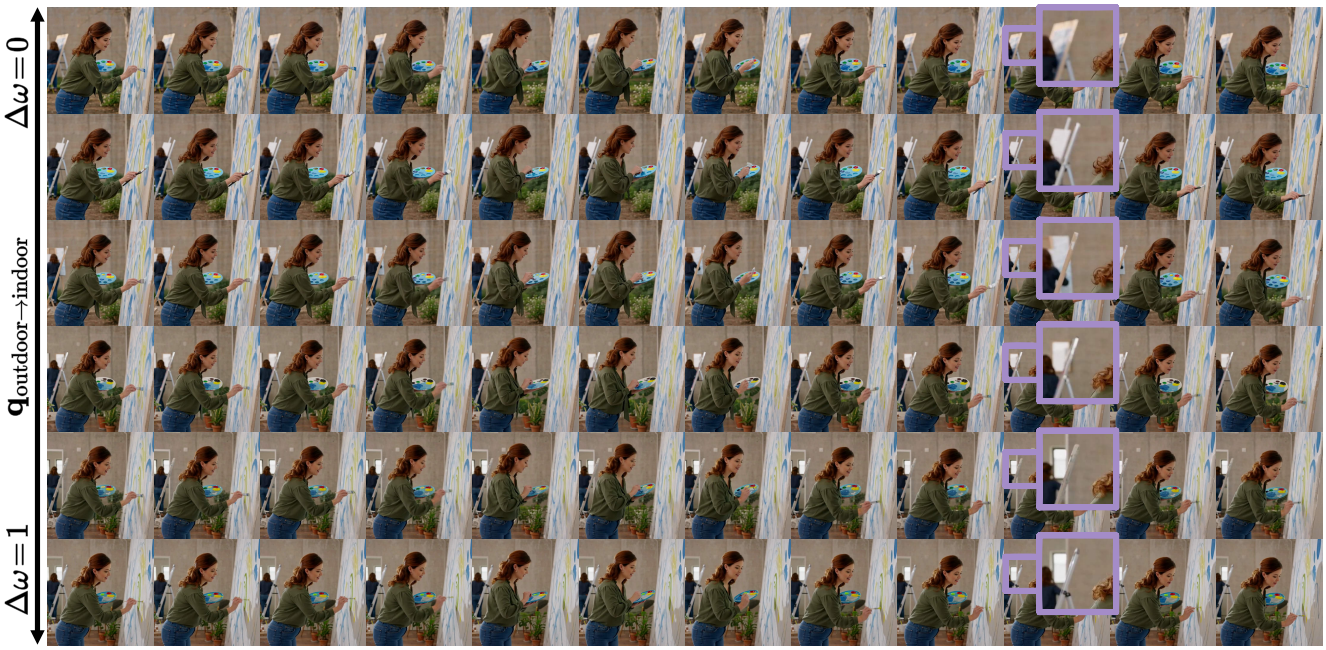
Figure 12. **Examples of action modulations**. Frame quality is compressed due to filesize (best viewed digitally).



(a) Video LLaMA 3 modulations between `parking lot` and `underpass`.

(b) Video LLaMA 3 modulations between `grass` and `sandtrap`.

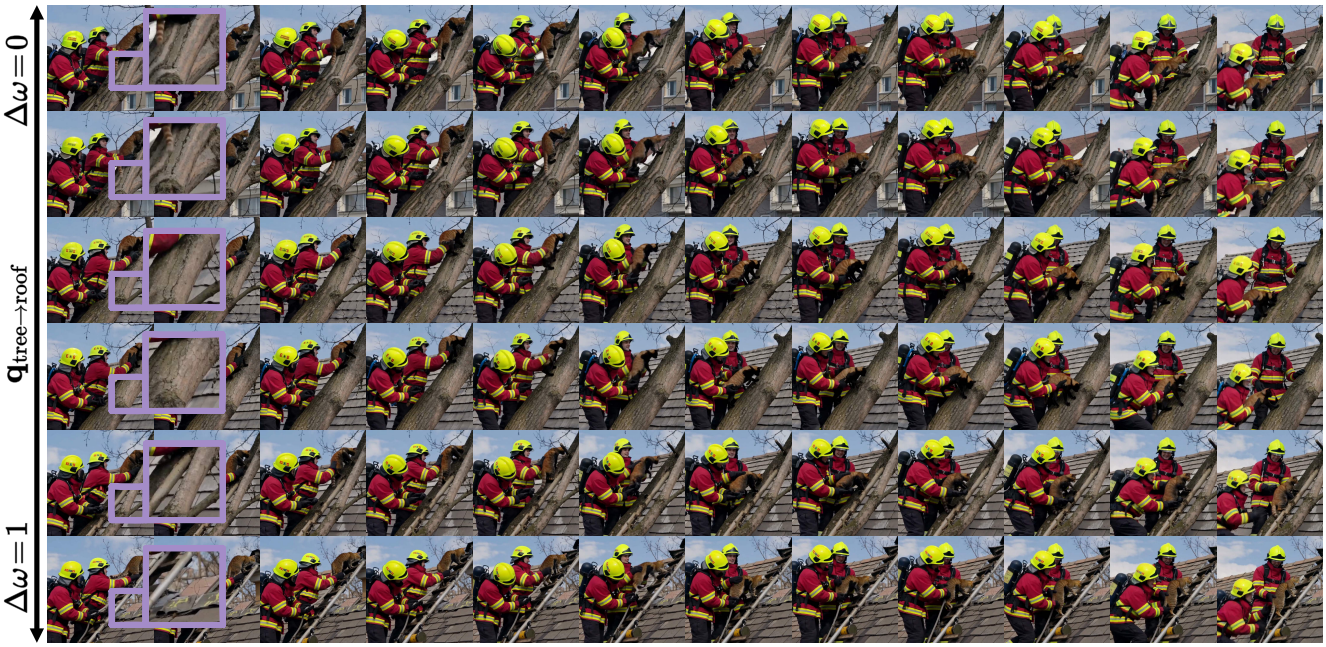
Figure 13. **Examples of scene modulations**. Frame quality is compressed due to filesize (best viewed digitally).



(c) Gemma 3 modulations between `outdoor` and `indoor`.

(d) Gemma 3 modulations between `clay` and `grass`.

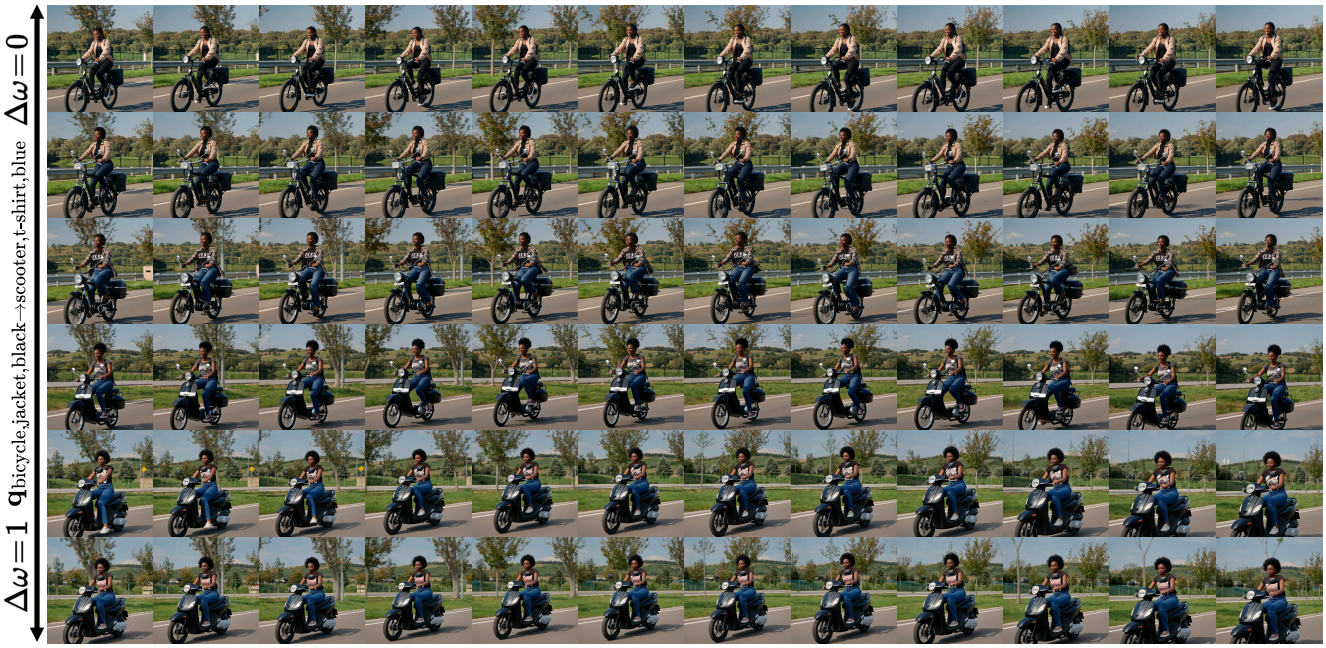
Figure 13. **Examples of scene modulations**. Frame quality is compressed due to filesize (best viewed digitally).



(e) Phi 4 MM modulations between `tree` and `roof`

(f) Phi 4 MM modulations between `overcast` and `clear sky`.

Figure 13. **Examples of scene modulations**. Frame quality is compressed due to filesize (best viewed digitally).



(a) Video LLaMA 3 modulations between `bicycle`, `jacket`, `black` and `scooter`, `t-shirt`, `blue`.

(b) Gemma 3 modulations between `black`, `night` and `red`, `morning`.

Figure 14. **Examples of multiple modulations**. Frame quality is compressed due to filesize (best viewed digitally).

(c) Phi 4 MM modulations between `canoeing`,`lake` and `kayaking`,`river`.

Figure 14. **Examples of multiple modulations**. Frame quality is compressed due to filesize (best viewed digitally).