LAVIB: A Large-scale Video Interpolation Benchmark Appendix

Table A1: **Vocabulary of search terms**. Terms are grouped to five main types including location, activities, weather, misc, and camera types. Search queries are the combination of multiple terms with an additional '4K'.

Location			Activities		Weather	Misc	Camera types
city	region	country	sports	actions			
[Amsterdam, Athens, Boston, Buenos Aires, Doha, Dubai, Istanbul, Lagos, Las Vegas, London, Los Angeles, Manchester, Mexico City, Miami, Montreal, New York, Paris, Perth, Porto, Rio de Janeiro, Seoul, Shanghai, Singapore, Tokyo, Venice, Vienna]	[Atlantic, California, Caribbean, England, Indian Ocean, Scandinavia, Sedona, Sicily, South East]	[Australia, Brazil, Bulgaria, Cambodia, Canada, China, Costa Rica, France, Germany, Iceland, India, Japan, Morocco, Mexico, Mongolia, Namibia, New Zealand, Nigeria, Russia, South Africa, Spain, Thailand]	[climbing, playing football, rafting, skating, skiing, snorkeling, snowboarding, tennis training]	[bike ride, car ride, dancing, exploration, walking]	[cloudy, overcast, rainy, snowing, sunny]	[Dolby Vision, PS5, animals, birds, flowers, forest, insects, marinelife, metro, mountains, ocean view, park, shoreline, car, underwater, wildlife, windmills]	[Blackmagic PCC 4K, Canon 5D Mark III, Canon EOS C200B, Canon EOS R6, DJI Inspire2, DJI OM4, DJI Osmo Pocket, GOPRO HERO10 Black, GOPRO HERO8 Black, GOPRO HERO9, GOPRO Max 360, Note 10 plus, RED RAVEN 4.5K, Samsung Galaxy, Sony A6700, Sony A7C, Yi 4K+, iPhone 12 Pro, iPhone 13 Pro]

A1 Vocabulary

Three core components are used for creating search terms from the vocabulary; locations, activities, or specific objects/settings relevant to videos. Locations and activities include two levels of hierarchies. The structure of search terms changes based on the selected sub-group.

A1.1 Location

Motivation. Natural scenes were found to have a large number of 4K footage from diverse camera types with minimal edits. Using an exhaustive list of locations is not feasible given the search space.

Remedy used. Instead, a list of locations was manually created based on the number of returned videos per location. Oversaturation of similar video locations; e.g. same country, was also manually adjusted for the selected terms.

About. The `city` subgroup is combined with a specific set of actions {`bike ride`, `car ride`, `exploration`, `walking`}. Weather conditions are added randomly to 1/3 of the search terms and camera types are added in 1/10, e.g.; 'Amsterdam bike ride rainy GOPRO HERO10 Black 4K'. It was seen that camera-type prompts can return results more relevant to the camera (e.g. reviews) and less relevant to the rest of the term searched. Thus, the probability of including camera types is kept low. For the `region` and `country` subgroups, prompts only include keywords such as 'best of' or 'scenic' as actions are less relevant when the locations are broad.

Limitations The manually-created list of locations does result in a level of selectivity. However, interpolation is a low-level computer vision task requiring only a basic understanding of scene dynamics and the general object shapes. Thus, the list's data diversity is believed to be sufficient. Tab. A2 reports results on the (full) LAVIB test set when training FLAVR on 700 videos from queries containing only either `London` , `Istanbul`, or `Seoul`.

Potential improvements. From Tab. A2, specific location terms do not show a significant impact on performance. However, including more locations can potentially further increase the variance of

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

some statistics; e.g. ARMS. In addition to weather queries, other terms such as time of day can be added to explicitly enforce diversification in the returned videos.

A1.2 Activities

Motivation. Activity terms are added to avoid static scenes. The distinction between sports and actions subgroups is done to control the expected motion intensity. Activities do however provide a strong constraint for the video content.

Remedy used. In total, approximately ~30% of the queries include actions. The majority of videos returned are either one-shot tours of locations or vlogs. Both types can easily be segmented into 10-second and 1-second clips by the pipeline as they include little to no edits/cuts. Sports are included in a small portion of the queries (4%) to avoid specialization. Tab. A3 ablates on 1,000 train videos sourced from queries that include different portions of activity terms. The evaluation is done on the (full) LAVIB test set. The partial inclusion of actions (30% and 60%) is shown to be the most balanced strategy for diversity.

About. Two activity categories are defined as motion variances, which present an important challenge in VFI. The `sports` subgroup primarily includes videos with fast-moving people/objects or camera motion. Specific terms are combined for the following sports; `climbing`, `rafting`, `skiing`, and `snowboarding` are combined with any of the {`forest`, `mountains`}, `snorkeling` is combined with {`marinelife`, `shoreline`, `underwater`}, and `tennis training` is combined with {`park`}. This results in search items such as; 'snowboarding mountain 4K'. The `action` subgroup is only used in combination with locations.

Limitations. Action terms such as `walking` or `car ride` are generic and return a large number of videos. Despite viewpoints from hand-held or mounted cameras being some of the most common in online videos, limitations exist.

Potential improvements. The videos returned using only location are primarily compilations/highlights from aerial, bird's eye, long shot, or panoramic footages. Driving videos are also ideal for capturing overhead shots. Although both help reduce viewpoint bias, adding an additional vocabulary term based on viewpoint can increase diversity further.

A1.3 Misc

Motivation. Miscellaneous search terms were manually added to diversify the search. The returned videos can vary from the rest of LAVIB by a. different luminance fluctuations; e.g. `underwater`, b. low; e.g. `metro` or c. high; e.g. `birds`, `flowers`, `insects`, contrast. 19 camera types are also selected manually to include a variety of phone cameras, action and digital cameras, and DSLRs. The difference in ARL and ALV distributions for misc and camera-based queries compared to the entire LAVIB is shown in Figs. A1 and A2.

Remedy used. Camera terms are added to ~10% of the queries to avoid returning irrelevant videos. This was done after manually checking the video titles. Approximately 7% of the dataset is collected with misc terms. Tab. A4 reports performance on 1,000 train examples that are partially sourced from misc queries. Similarly to Tab. A3, maintaining a balance between misc and non-misc queries improves generalizability.

About. Miscellaneous search terms are primarily combined with recording equipment to form queries; e.g. 'ocean view Yi 4K+ 4K'.

Limitations. The inclusion of misc terms aims to improve diversity. However, as noted, video themes such as screen captures do not guarantee significant variations in video statistics. The narrow ALV/ARL distributions of videos sourced from misc queries are compared to an equally sized random sample from LAVIB in Fig. A3. Similarly, some camera types may not necessarily differ in video quality.

Potential improvements. A further analysis on the misc queries that source videos with the most diverse statistics can highlight the specific terms that improve variance. This can also be used to weigh each term during selection. The same approach can also be applied to the camera types.

A2 Video sorting

For the benchmark, each split should have similar video metric distributions. Due to the multi-dimensionality and high variance across metrics, DUPLEX is used to calibrate dataset split sampling.

Table A2: **Results on different location-based train subsets.**

Term	PSNR↑	SSIM↑
London	30.69	0.945
Istanbul	30.75	0.944
Seoul	**30.81**	**0.949**

Table A3: **Results on different activity-based subsets.**

Act. (%)	PSNR↑	SSIM↑
0	28.57	0.932
30	**31.08**	**0.953**
60	30.65	0.948
100	29.23	0.940

Table A4: **Results on different misc-based subsets.**

Misc (%)	PSNR↑	SSIM↑
0	29.31	0.941
30	**30.54**	**0.950**
60	29.66	0.934
100	28.83	0.929

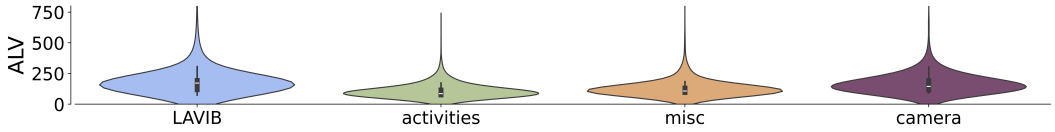


Figure A1: **ALV distributions** for all LAVIB and videos from activities, misc, and camera queries.

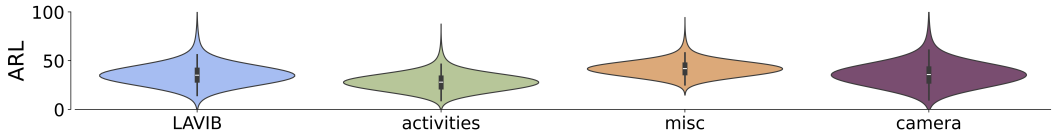
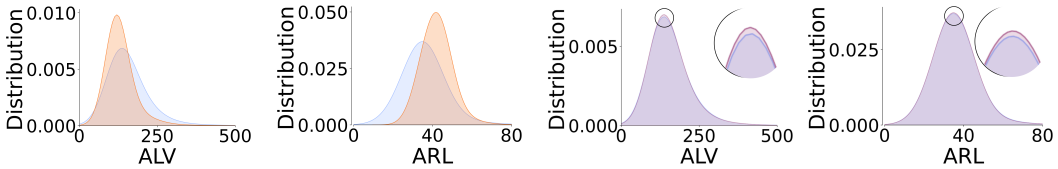


Figure A2: **ARL distributions** for all LAVIB and videos from activities, misc, and camera queries.



(a) ALV of all LAVIB and misc-only sourced videos. (b) ARL of all LAVIB and misc-only sourced videos. (c) ALV of all LAVIB and a random subset of videos. (d) ARL of all LAVIB and a random subset of videos.

Figure A3: **ARL and ALV distributions** for all LAVIB videos. Additional distributions from misc queries and a random subset, both of size 19,706 (~7% of total videos), are shown for direct comparisons.

DUPLEX uses the L2 distance across video metrics when creating train/val/test splits. For each set, the algorithm discovers the two most distant videos given their AFM, ALV, ARMS, and ARL metrics. It then iteratively samples videos that maximize the distance to previously sampled videos. This is done iteratively until the size condition for the split is met. Algorithm 1 provides a programmatic view of DUPLEX sampling.

A3 Detailed training settings

All training experiments are done with the codebases provided by the authors with 2× Nvidia L40 with an average training time of 2 days per model. Computational settings for each model are reported in Tabs. A5 to A7.

A4 Ablations

Supplementary to the main results in §5 ablations are performed with FLAVR for variations in train set sizes for both benchmark and OOD challenges.

Benchmarks over reduced training set sizes. Tab. A8 presents val and test set results with reductions in the training set sizes. At each reduction setting, clips are dropped randomly. Performance drops significantly for both validation and test sets as the size of the training set decreases with an average -4.12 and -0.086 PSNR/SSIM.

Performance over varying size. Motivated by the performance reductions observed with decreases in the train set size in Tab. A8, Fig. A4 presents PSNR/SSIM performance when an additional number of clips is retained during the selection progress. Clips are added by relaxing the threshold values.

Algorithm 1 DUPLEX video selection

Input: dataset $\mathcal{D}$, sets { train, val, test }
Output: dataset splits: $\{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}\}$

for set_i $\in \{\text{train}, \text{val}, \text{test}\}$ **do**
  $\mathbf{s} \leftarrow \max(\{||\mathbf{x}_j - \mathbf{x}_k^T||_2\})$ where $\mathbf{x}_j, \mathbf{x}_k$ are metrics for videos j, k within $\mathcal{D}$.
  $\mathcal{D}_{\text{set\_i}} \leftarrow \{j, k\}$
  $\mathcal{D} \leftarrow \mathcal{D} \setminus \{j, k\}$
end for
for set_idx $\in \{tr, v, ts\}$ **do**
  for $i \in \{3, \text{set\_idx}\}$ **do**
    $\mathbf{s}_i \leftarrow \max(\{||\mathbf{x}_l - \mathcal{D}_{\text{set\_idx}}[-1]^T||_2\})$ where $\mathbf{x}_l$ is a video in $\mathcal{D}$.
    $\mathcal{D}_{\text{set\_idx}} \leftarrow \mathcal{D}_{\text{set\_idx}} \cup \{\mathbf{x}_l\}$
    $\mathcal{D} \leftarrow \mathcal{D} \setminus \{\mathbf{x}_l\}$
  end for
end for

Table A5: **RIFE settings**

Parameter	value
batch size	64
optimizer	AdamW
weight decay	$1e^{-6}$
learning rate	$1e^{-4}$
learning scheduler	Step
additional params	beta1=0.9 beta2=0.99

Table A6: **EMA-VFI settings**.

Parameter	value
batch size	64
optimizer	AdamW
weight decay	$1e^{-4}$
learning rate	$1e^{-4}$
learning scheduler	Warmup
additional params	beta1=0.9 beta2=0.99

Table A7: **FLAVR settings**

Parameter	value
batch size	64
optimizer	Adam
weight decay	$1e^{-6}$
learning rate	$5e^{-3}$
learning scheduler	Step
additional params	beta1=0.9 beta2=0.99

Table A8: **Val and test set results** when training on different portions of the train set. *full* denotes that the entire train set from LAVIB is retained for training. Best results per split are in **bold**.

LAVIB train %	val set PSNR↑	val set SSIM↑	val set LPIPS↓	test set PSNR↑	test set SSIM↑	test set LPIPS↓
20%	29.43	0.895	$8.257e^{-2}$	29.48	0.894	$8.472e^{-2}$
40%	31.68	0.960	$4.108e^{-2}$	31.63	0.958	$4.241e^{-2}$
60%	32.64	0.970	$3.566e^{-2}$	32.50	0.967	$3.835e^{-2}$
80%	33.36	0.975	$2.971e^{-2}$	33.19	0.973	$3.064e^{-2}$
full	**33.72**	**0.981**	**$2.515e^{-2}$**	**33.44**	**0.981**	**$2.934e^{-2}$**

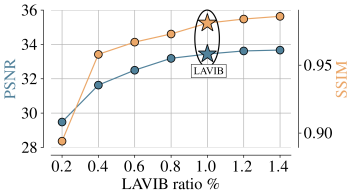


Figure A4: **Test PSNR/SSIM over train sizes**. In ratios < 1.0% clips are removed. In ratios > 1.0% clips are added from left-out segments.

Although the performance improvements observed when including more clips in training in small ratios are significant, this is not retraced with further increases in the size of the current dataset. This shows that the selection process for LAVIB enables the creation of a diverse dataset.

OOD over reduced train set sizes. Performance trends when removing the highest/lowest valued clips in OOD challenges given their metrics are reported in Tab. A9 and Tab. A10 for AFM and ALV. Similarly, Tab. A11 and Tab. A12 report results with training set reductions for ARMS and ARL. Across settings, the portions closer to the target domain; e.g. the top 30% for the low to high settings and bottom 30% for the high to low settings present the largest drop in performance when removed. In contrast, when portions of the data that are less similar to the target domain are removed the reductions in performance are marginal. This shows that VFI method trained on domain-specific videos cannot generalize as effectively.

A5 Qualitative

Fig. A5 presents predicted frames from each model on examples from the benchmark test set. Interpolated frames for AFM-, ALV-, ARMS-, and ARL-based OOD challenges are shown in

Table A9: **AFM OOD ablation results**.

AFM sett.	Train % removed	PSNR↑	SSIM↑	LPIPS↓
	- bottom 30%	29.45	0.912	$5.637e^{-2}$
	- top 30 %	26.32	0.854	$9.428e^{-2}$
	None	**30.67**	**0.959**	**$5.094e^{-2}$**
	- bottom 30%	32.80	0.973	$3.396e^{-2}$
	- top 30 %	35.32	0.987	$1.503e^{-2}$
	None	**35.66**	**0.991**	**$1.342e^{-2}$**

Table A10: **ALV OOD ablation results**.

ALV sett.	Train % removed	PSNR↑	SSIM↑	LPIPS↓
	- bottom 30%	30.32	0.933	$4.781e^{-2}$
	- top 30 %	28.54	0.905	$5.567e^{-2}$
	None	**31.78**	**0.962**	**$2.942e^{-2}$**
	- bottom 30%	31.24	0.958	$4.396e^{-2}$
	- top 30 %	34.35	0.971	$2.740e^{-2}$
	None	**34.67**	**0.975**	**$2.627e^{-2}$**

Table A11: **ARMS OOD ablation results**.

ARMS sett.	Train % removed	PSNR↑	SSIM↑	LPIPS↓
	- bottom 30%	32.87	0.977	$2.683e^{-2}$
	- top 30 %	31.92	0.965	$3.769e^{-2}$
	None	**33.02**	**0.982**	**$2.561e^{-2}$**
	- bottom 30%	30.18	0.931	$4.515e^{-2}$
	- top 30 %	30.74	0.973	$3.327e^{-2}$
	None	**31.11**	**0.977**	**$3.024e^{-2}$**

Table A12: **ARL OOD ablation results**.

ARL sett.	Train % removed	PSNR↑	SSIM↑	LPIPS↓
	- bottom 30%	32.76	0.973	$2.806e^{-2}$
	- top 30 %	32.15	0.961	$3.315e^{-2}$
	None	**33.97**	**0.980**	**$2.543e^{-2}$**
	- bottom 30%	34.06	0.972	$2.763e^{-2}$
	- top 30 %	33.67	0.970	$3.457e^{-2}$
	None	**34.20**	**0.976**	**$2.875e^{-2}$**

Figs. A6 to A9. In all settings, models can only partially interpolate the unseen frames. The majority of the errors observed are related to high-motion low-contrast examples. In instances where motion blur is present in the ground truth; e.g row 2 Fig. A5, row 5 in Fig. A6, and rows 1,5, and 6 in Fig. A7, motion blur is exacerbated at the interpolated frames from all models. Models trained on settings where fine details are not visible such as low ARMS and low ARL only interpolate the general shapes of objects and structures as shown in rows 1 and 2 in Fig. A8 and rows 1–3 in Fig. A9.

A6 Ethics, privacy, and use

Ethics and privacy. The introduced dataset primarily considers footage of landscapes, objects, nature, animals, and screen recordings. However, certain videos may include people. Scenes in which people appear are characterized by high camera motion, scene clutter, and partial visibility of faces that appear briefly for a few seconds. Thus, it is believed that the risk of identification is low. In addition, the video segments from which the dataset is sourced are 1 second long, limiting the number of frames available. As videos are sourced from YouTube a list of the links to the original videos is also provided.

Use. The dataset is distributed for open-source scientific projects under a Creative Common's Attribution-NonCommercial-Share-Alike (CC BY-SA-NC 4.0). The dataset can be further shared, and adapted, but cannot be used for commercial applications. Adaptations or sharing of the dataset needs to be done under the same license[†].

[†]Clarifications on special use cases can be found in: https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

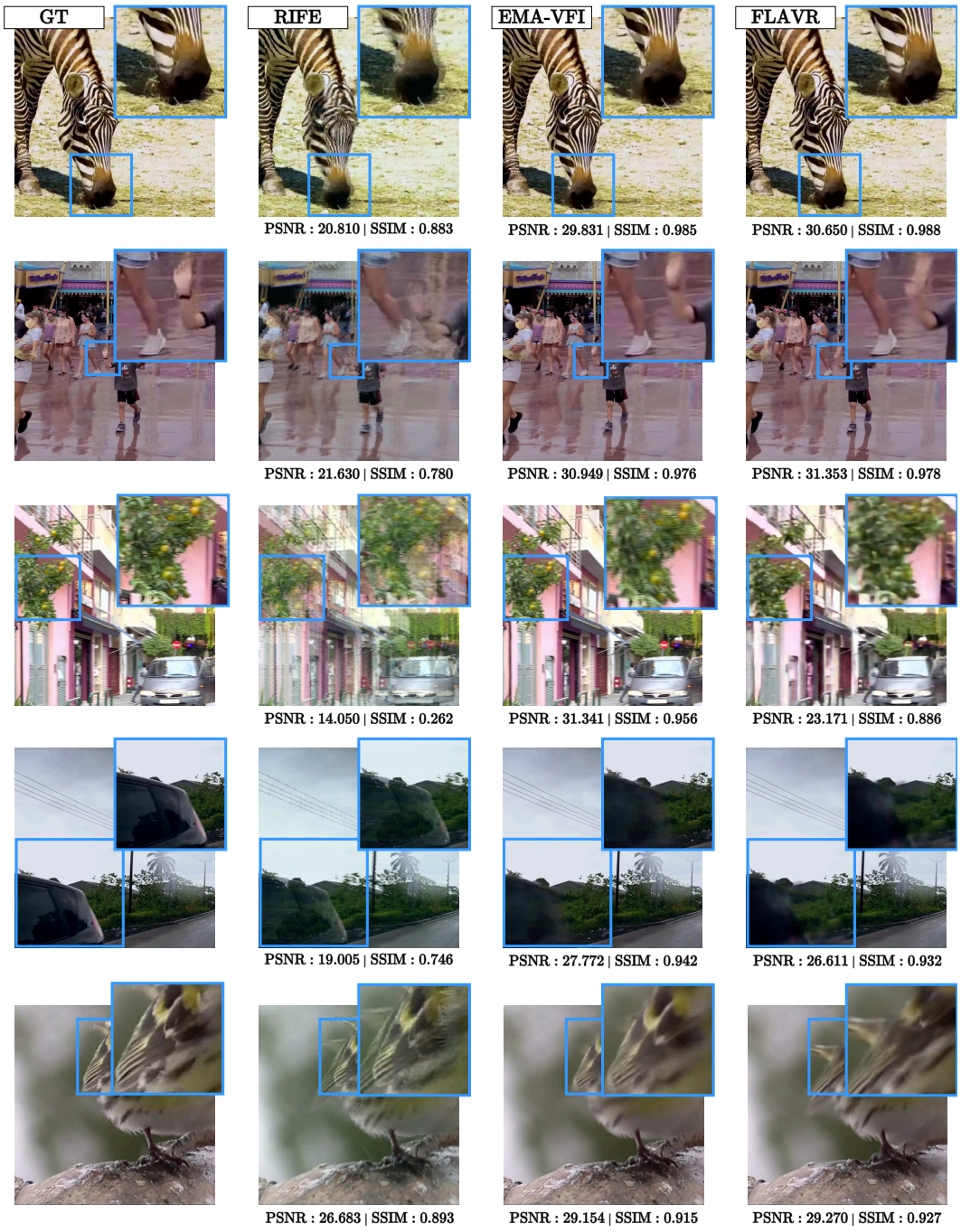


Figure A5: **Examples from the LAVIB benchmark** (best viewed digitally)

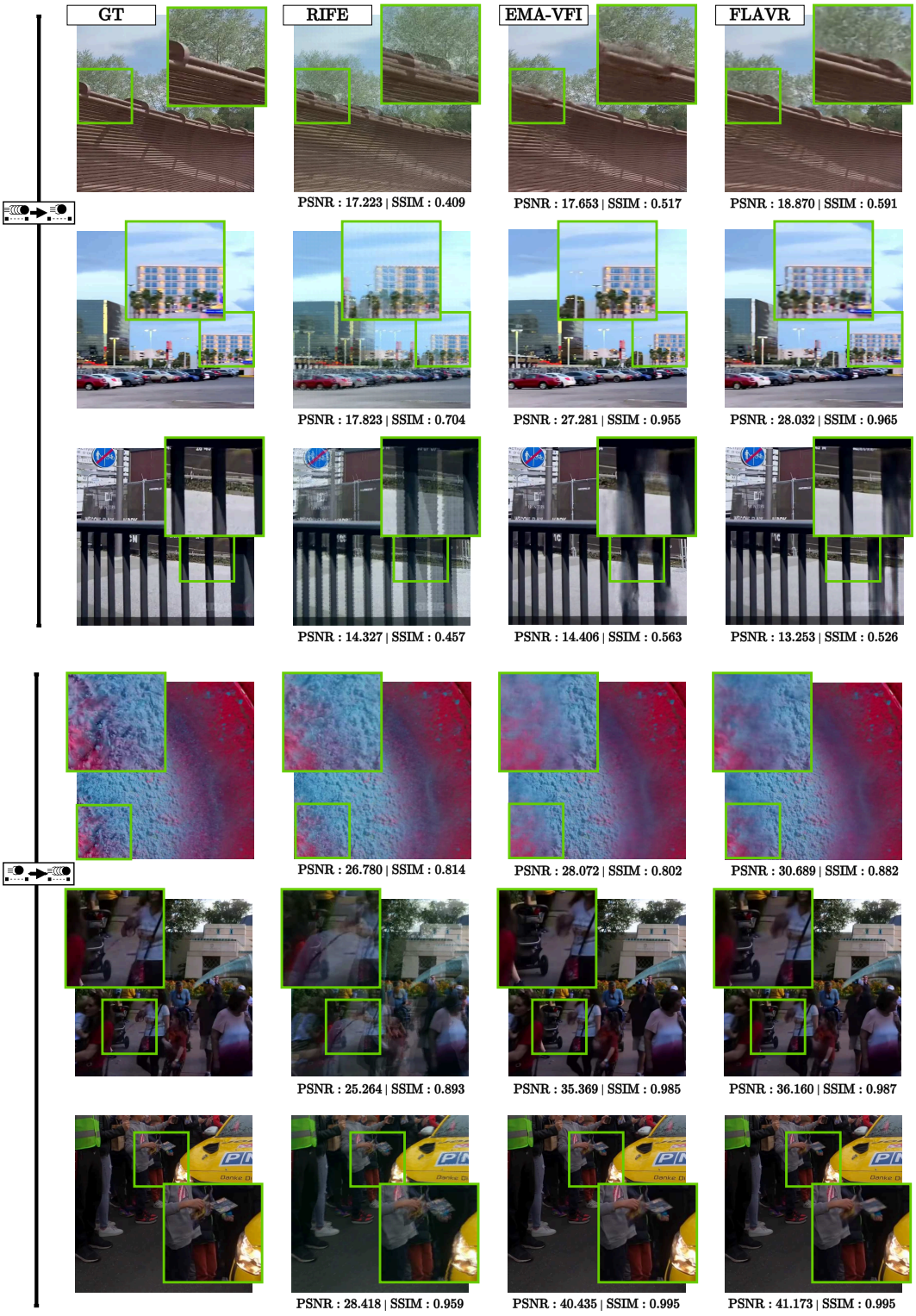


Figure A6: **Examples of AFM OOD challenges** (best viewed digitally)

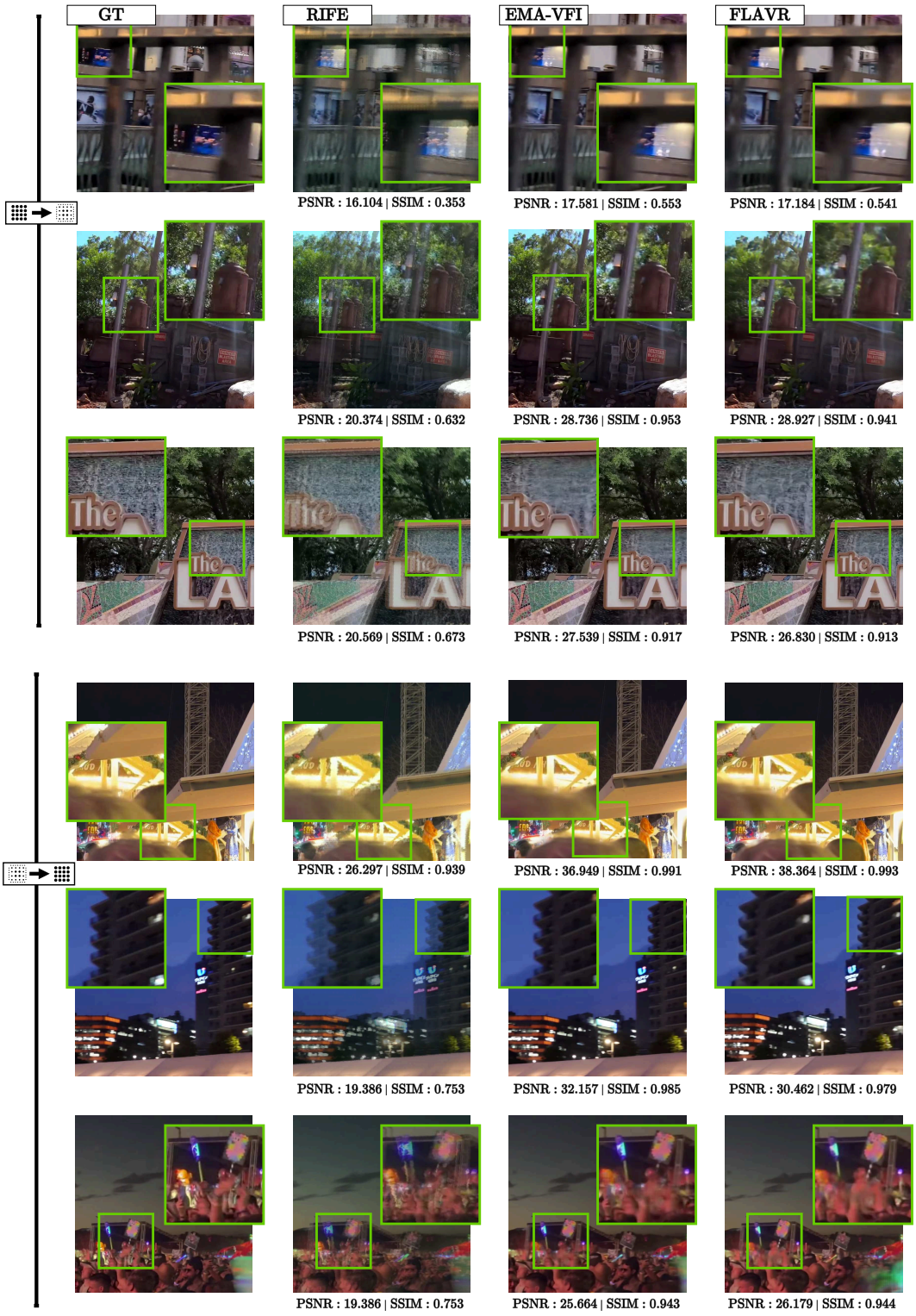


Figure A7: **Examples of ALV OOD challenges** (best viewed digitally)

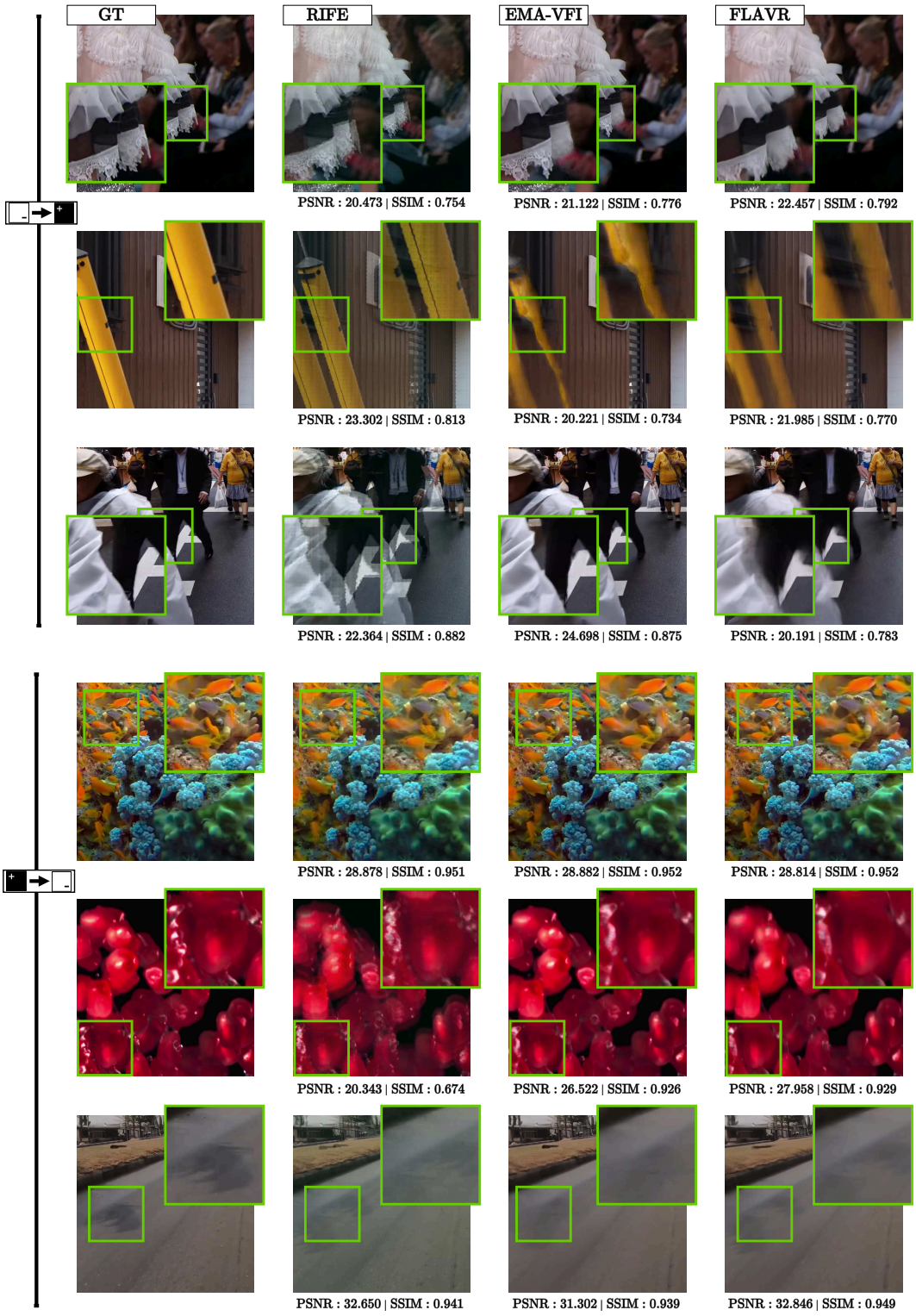


Figure A8: **Examples of ARMS OOD challenges** (best viewed digitally)

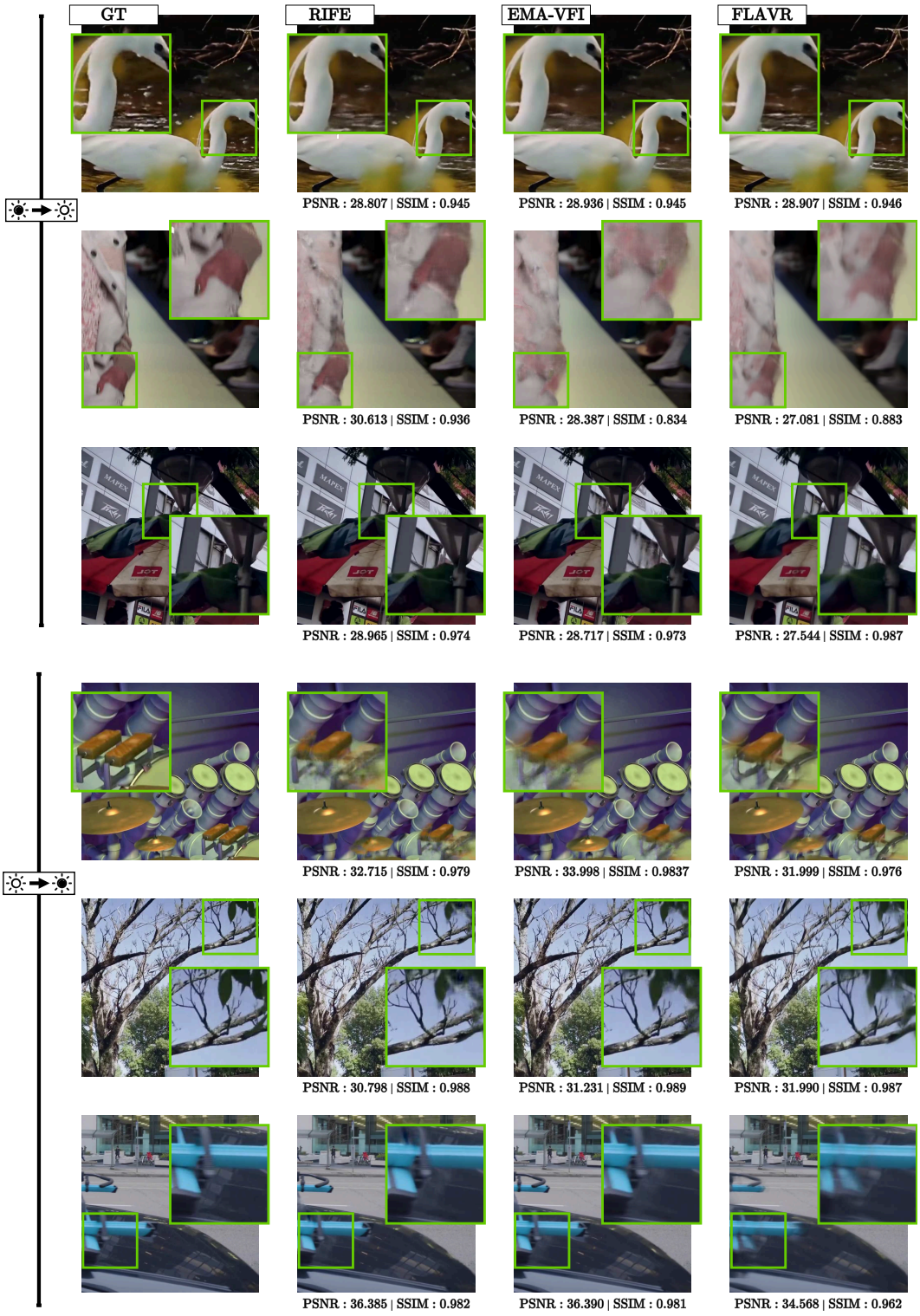


Figure A9: **Examples of ARL OOD challenges** (best viewed digitally)