VGA-Bench: A Unified Benchmark and Multi-Model Framework for Video Aesthetics and Generation Quality Evaluation

Supplementary Material

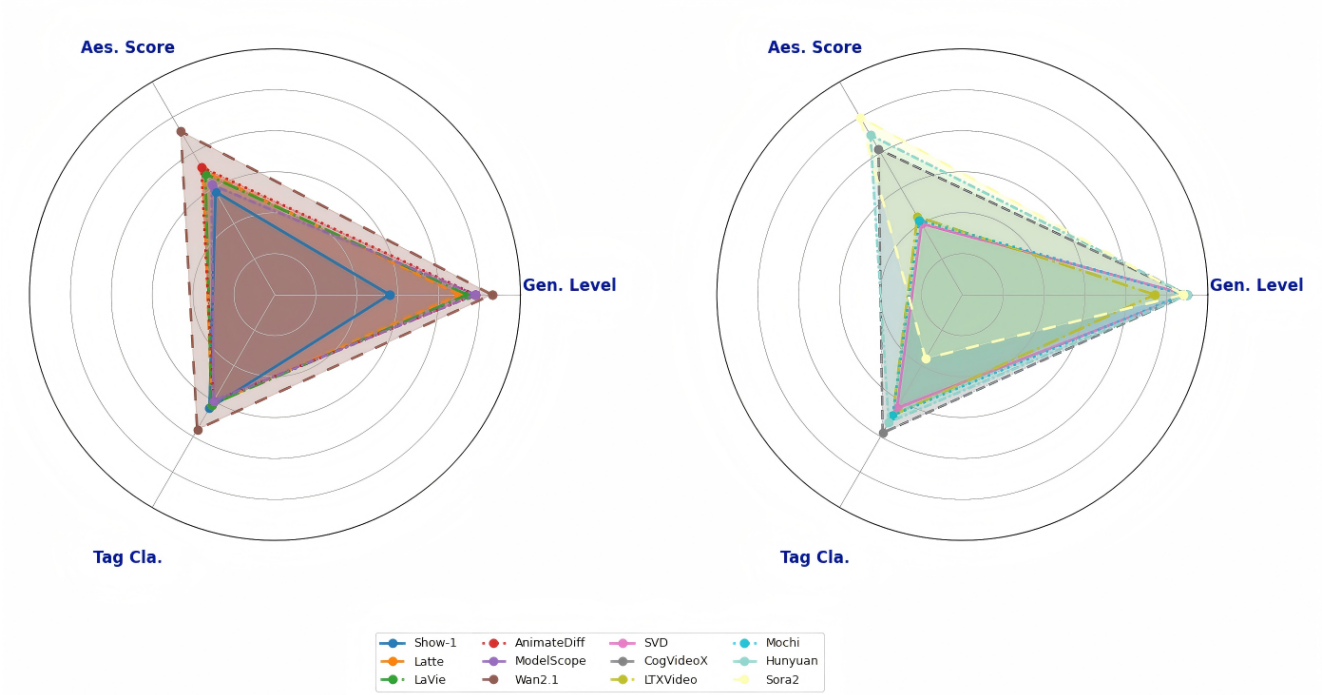

Figure 1. Radar chart comparing the performance of various video generation models across three evaluation dimensions. The concentric circular grids represent score levels, ranging from 0.1 at the center to 0.9 outward in 0.1 intervals. Higher values indicate better performance. Each model is represented by a closed polygon with distinct colors and line styles for easy comparison.

1. Dimension of Generation Quality

The meanings of the dimensions of generation quality are summarized in Table 3.

2. Overview of Generative Model Performance

The performance comparison of various generative models across the three core dimensions is shown in Figure 1.

3. Comprehensive Evaluation Results of Generative Models across Sub-Dimensions in VGA-Bench

3.1. Performance Comparison of Generative Models on Aesthetic Quality Dimensions

Table 1 presents the scoring results of 12 mainstream generative models across sub-attributes under the aesthetic quality dimension. All scores are generated by VAQA-Net through automated evaluation, reflecting the visual appeal and artistic expressiveness of the videos produced by these models.

3.2. Comparison of Aesthetic Tag Prediction Capabilities

Table 2 reports aesthetic tag prediction accuracies using VTag-Net, evaluating the models' capabilities in understanding and generating complex aesthetic semantics.

3.3. Performance Comparison of Generative Models on the Generation Quality Dimension

Due to the extensive sub-dimensions within generation quality, we conduct automated annotations using VGQA-Net and present the performance comparisons across three separate tables: Table 4 details 11 metrics on video-text consistency and spatio-temporal alignment; Table 5 assesses 14 realism metrics concerning physical laws and real-world commonsense; and Table 6 evaluates 6 technical dimensions reflecting basic low-level visual fidelity.

Table 1. Performance Comparison of Generative Models on Aesthetic Quality Dimensions

Model	Overall	Com	SS	Lig	VT	Col	DoF	Exp	Cos	Mak
Show-1 [**?**]	0.290	0.415	0.383	0.367	0.357	0.397	0.330	0.294	0.317	0.294
Latte-1 [**?**]	0.345	0.472	0.435	0.420	0.410	0.445	0.393	0.313	0.340	0.296
LaVie [**?**]	0.341	0.468	0.435	0.421	0.407	0.442	0.396	0.309	0.340	0.293
AnimateDiff [**?**]	0.356	0.475	0.446	0.427	0.415	0.460	0.400	0.317	0.364	0.310
ModelScope [**?**]	0.312	0.445	0.413	0.397	0.381	0.430	0.377	0.287	0.334	0.266
Wan2.1 [**?**]	0.459	0.552	0.523	0.523	0.521	0.521	0.503	0.436	0.459	0.412
SVD [**?**]	0.204	0.305	0.260	0.229	0.233	0.277	0.191	0.207	0.151	0.164
CogVideoX [**?**]	0.405	0.480	0.454	0.437	0.431	0.453	0.399	0.389	0.420	0.354
LTXVideo [**?**]	0.214	0.313	0.275	0.238	0.241	0.285	0.208	0.189	0.142	0.125
Mochi [**?**]	0.211	0.306	0.269	0.240	0.234	0.283	0.203	0.198	0.147	0.147
Hunyuan [**?**]	0.452	0.531	0.505	0.507	0.504	0.512	0.485	0.455	0.465	0.430
Sora2 [**?**]	**0.504**	**0.559**	**0.532**	**0.543**	**0.564**	**0.560**	**0.540**	**0.520**	**0.532**	**0.452**

Table 2. Comparison of Aesthetic Tag Prediction Capabilities

Model	CT	NoLS	LSP	LQ	LC	ST	DoF	Sat	Bri	Col	Con
Show-1[**?**]	0.17	0.47	0.19	0.46	0.17	0.32	0.54	0.29	0.42	0.48	0.35
Latte-1[**?**]	0.16	0.47	0.19	0.46	0.12	0.25	0.60	0.29	0.35	0.46	0.35
LaVie[**?**]	0.17	0.47	0.19	0.46	0.11	0.25	0.66	0.29	0.41	0.39	0.35
AnimateDiff[**?**]	0.18	0.47	0.19	0.46	0.06	0.23	0.66	0.29	0.37	0.39	0.35
ModelScope[**?**]	0.17	0.47	0.19	0.46	0.01	0.23	0.60	0.29	0.43	0.40	0.35
Wan2.1[**?**]	0.18	0.60	0.19	**0.57**	0.23	0.34	**0.68**	**0.38**	0.47	0.58	0.36
SVD[**?**]	**0.19**	0.50	0.24	0.46	0.28	0.22	0.52	0.29	0.25	0.46	**0.37**
CogVideoX[**?**]	0.18	**0.71**	0.19	0.51	0.33	0.38	0.62	0.29	**0.48**	0.63	**0.37**
LTXVideo[**?**]	0.15	0.52	**0.27**	0.46	0.36	0.20	0.57	0.29	0.25	0.64	0.36
Mochi[**?**]	0.17	0.49	0.24	0.46	**0.44**	0.20	0.55	0.29	0.25	**0.69**	0.35
Hunyuan[**?**]	0.16	0.66	0.21	0.48	0.23	**0.40**	**0.68**	0.27	0.45	0.47	0.35
Sora2[**?**]	0.13	0.32	0.14	0.29	0.19	0.32	0.27	0.11	0.15	0.19	0.10

Table 3. Number and explanation of different assessment dimensions

Type	Num.	Assessment Dimension	Description
Video-Text Consistency	1	Character-Text Consistency	Whether specific characters in the video match the text description (e.g., Elon Musk should appear as the correct individual).
	2	Action-Text Consistency	Whether actions in the video match the text description (e.g., running, jumping), focusing solely on the action regardless of the subject.
	3	Scene-Text Consistency	Whether scenes in the video match the described settings (e.g., hospital, school), including identifiable scene elements.
	4	Object Position-Text Consistency	Object positions refer to relative placement based on camera orientation (e.g., if "a motorcycle is to the left of a bus," they should appear on corresponding sides of the video frame).
	5	Object Attribute-Text Consistency	Object attributes include descriptive features like color, shape, and texture.
	6	Object-Text Consistency	Whether objects in the video can be correctly identified as those mentioned in the text.
	7	Video Content-Text Consistency	Overall alignment where every textual description should be accurately generated.
	8	Video Speed-Text Consistency	Whether video speed matches textual descriptions (current samples only include slow-motion).
	9	Video Style-Text Consistency	Whether artistic styles mentioned in text (e.g., Van Gogh, Picasso) are recognizable in the video.
	10	Camera Movement-Text Consistency	Whether camera movements described in text (e.g., pan left, tilt right) are properly executed.
	11	Unrealistic Description Imaginative Presentation	When text describes unrealistic scenarios (e.g., "an astronaut riding a horse in space"), whether the video presentation aligns with imaginative expectations.
Realism & Plausibility	12	Rigid Body Collision Realism	Whether rigid body collisions in videos appear physically plausible.
	13	Action Realism	Whether actions could realistically be performed.
	14	Scene Realism	Whether scenes appear sufficiently realistic when no special style is specified in text.
	15	Weather Representation Realism	Whether weather conditions appear realistic.
	16	Time Period Representation Realism	Whether time-period representations appear authentic.
	17	Gaseous Motion Realism	Whether gas dynamics (smoke, vapor) appear physically accurate.
	18	Fluid Motion Realism	Whether fluid movements appear physically plausible.
	19	Gradual Change Motion Realism	Whether gradual transformations (balloon inflation, plant growth) appear physically accurate.
	20	Object Motion Trajectory Realism	Whether object movement paths follow physically plausible dynamics.
	21	Object Realism	Whether objects appear sufficiently realistic.
	22	Character Generation Quality	Whether human characters appear sufficiently realistic.
	23	Textual Attribute Representation Realism	Whether object attributes (color, shape, texture) match real-world appearances.
	24	Video Lighting and SGQAow Realism	Whether lighting and sGQAows appear physically accurate.
	25	Moving Scene Reasonableness	Whether scene transitions during camera movements maintain proper perspective.
	26	Overall Realism	Whether the entire video looks realistic overall.
Basic Quality	27	Abnormal Lighting Detection	Videos should avoid lighting artifacts (overexposure, abnormal flares).
	28	Video Noise-Free	Videos should exhibit no noticeable noise artifacts.
	29	Video Clarity	Whether video resolution is sufficiently sharp.
	30	Static Content Non-distortion	Stationary objects shouldn't distort abnormally during camera movement.
	31	Static Content Stability	Stationary objects shouldn't distort abnormally over time (temporal consistency).

Table 4. Evaluation Results on Video-Text Consistency

Model	1	2	3	4	5	6	7	8	9	10	11
Show-1[**?**]	0.037	0.364	0.362	0.346	0.505	0.466	0.196	0.027	0.030	0.433	0.250
Latte-1[**?**]	0.052	0.285	0.770	0.565	0.875	0.672	0.598	0.000	0.034	0.500	0.288
LaVie[**?**]	**0.076**	0.212	0.852	0.512	0.887	0.682	0.598	0.007	0.064	0.612	0.340
AnimateDiff[**?**]	0.024	0.168	0.783	0.500	**0.977**	0.770	0.547	0.000	0.014	**0.861**	0.400
ModelScope[**?**]	0.021	0.208	0.754	0.464	0.964	0.816	0.567	0.014	0.037	0.806	**0.480**
Wan2.1[**?**]	0.026	0.840	0.672	0.629	0.900	0.795	0.628	**0.295**	**0.184**	0.595	0.448
SVD[**?**]	0.066	**0.958**	**0.893**	0.572	0.934	**0.917**	0.591	0.071	0.050	0.655	0.410
CogVideoX[**?**]	0.055	0.936	0.863	0.512	0.929	0.888	0.599	0.045	0.065	0.720	0.419
LTXVideo[**?**]	0.032	0.561	0.578	0.542	0.781	0.640	0.538	0.038	0.043	0.533	0.303
Mochi[**?**]	0.039	0.897	0.803	**0.701**	0.951	0.813	**0.679**	0.039	0.043	0.632	0.387
Hunyuan[**?**]	0.056	0.856	0.799	0.592	0.970	0.809	0.600	0.034	0.041	0.620	0.350
Sora2[**?**]	0.024	0.842	0.790	0.532	0.934	0.851	0.603	0.086	0.094	0.715	0.425

Table 5. Evaluation Results on Realism & Plausibility

Model	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Show-1[**?**]	0.028	0.267	0.294	0.377	0.206	0.344	0.377	0.609	0.127	0.351	0.237	0.389	0.284	0.282	0.227
Latte-1[**?**]	0.038	0.435	0.511	0.640	0.388	0.350	0.488	0.825	0.150	0.428	0.513	0.600	0.413	0.463	0.598
LaVie[**?**]	0.035	0.402	0.529	0.618	0.345	0.415	0.490	0.830	0.145	0.408	0.530	0.607	0.440	0.533	0.598
AnimateDiff[**?**]	0.025	0.490	0.535	0.644	0.415	0.390	0.500	0.760	0.145	0.466	0.611	0.677	0.485	0.556	0.547
ModelScope[**?**]	0.070	0.484	0.522	0.676	0.405	0.375	0.490	0.870	0.180	0.508	0.574	0.660	0.490	0.574	0.567
Wan2.1[**?**]	**0.163**	0.517	0.521	0.539	**0.481**	0.575	0.500	**0.938**	**0.210**	0.527	0.521	0.572	**0.494**	0.533	0.628
SVD[**?**]	0.025	0.496	**0.586**	0.672	0.445	**0.655**	0.500	0.660	0.165	0.605	0.652	0.640	0.440	**0.671**	0.591
CogVideoX[**?**]	0.030	0.512	0.551	0.677	0.461	0.597	0.470	0.875	0.139	0.590	0.631	0.602	0.420	0.641	0.599
LTXVideo[**?**]	0.035	0.465	0.520	0.605	0.391	0.471	0.444	0.864	0.131	0.497	0.520	0.595	0.424	0.509	0.623
Mochi[**?**]	0.047	0.506	0.540	0.665	0.415	0.500	**0.560**	0.937	0.174	0.535	0.653	0.631	0.466	0.578	**0.786**
Hunyuan[**?**]	0.035	0.526	0.563	**0.784**	0.428	**0.526**	0.547	0.864	0.172	0.593	**0.714**	**0.700**	0.438	0.616	0.694
Sora2[**?**]	0.085	0.520	0.553	0.660	0.450	0.585	0.490	0.780	0.190	**0.620**	0.604	0.694	0.450	0.668	0.603

Table 6. Evaluation Results on Basic Visual Quality

Model	27	28	29	30	31
Show-1[**?**]	0.369	0.097	0.212	0.257	0.285
Latte-1[**?**]	0.766	0.158	0.365	0.447	0.742
LaVie[**?**]	0.861	0.207	0.452	0.500	0.717
AnimateDiff[**?**]	0.843	0.214	0.526	0.525	0.680
ModelScope[**?**]	0.633	0.232	0.454	0.482	0.664
SVD[**?**]	0.898	0.292	**0.650**	0.511	0.733
CogVideoX[**?**]	0.855	**0.304**	0.592	0.515	0.748
LTXVideo[**?**]	0.823	0.238	0.476	0.518	**0.808**
Mochi[**?**]	0.824	0.223	0.500	0.511	0.805
Hunyuan[**?**]	0.876	0.253	0.556	**0.578**	0.797
Sora2[**?**]	**0.906**	0.215	0.540	0.518	0.720