This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

Research Article

https://doi.org/10.26599/OCEAN.2026.9470033

Prompt engineering for synthetic deep-sea mining video: A controlled evaluation of four text-to-video systems

Nan Li[1], Long Yu[2,*], Chaojun Huang[3], Yi Yang[3], Kanglin Liu[1]

1 Beijing Jiaotong University, Beijing 100044, China

2 Tsinghua University, Beijing 100084, China

3 China Merchants Marine and Offshore Research Institute Co., Ltd., Shenzhen 518000, Guangdong, China

Address correspondence to Long Yu, yulong@mail.tsinghua.edu.cn

Received: July 13, 2026 / Revised: September 25, 2026 / Accepted: October 13, 2026

Abstract. Deep-sea mining video is scarce, and low illumination, sediment plumes, and moving equipment challenge visual monitoring. We evaluated four text-to-video systems on six VBench dimensions, first in an archived 300-video system-level comparison and then in a 240-video extension crossing four prompt conditions, five scenarios, and three requested seeds. Within-system prompt contrasts varied by metric; detailed or negatively constrained prompts did not consistently raise VBench scores. All 15 Wan pairs generated with and without a negative-prompt request had pixel-identical decoded frames, leaving that contrast uninformative for this service. In a separate 60-video review with system identities masked, three raters' majority labels identified at least one domain violation in 37 clips (61.7%; descriptive Wilson 95% interval, 49.0–72.9%); agreement differed by criterion. These results support VBench scoring and domain-rule review as complementary first-stage screens, while physical fidelity and downstream utility remain to be tested.

Keywords: Video generation models; Prompt engineering; Deep-sea mining; Model evaluation benchmarks

1 Introduction

Deep-sea mining has attracted increasing attention as a means of exploiting polymetallic nodules and other seabed mineral resources, although its environmental impacts and associated knowledge gaps remain under debate [1]. During mining operations, seabed vehicles, collection mechanisms, soft sediments, suspended particles, and discharged tailings interact in a highly coupled environment [2, 3]. These interactions may cause crawler sinkage, sediment resuspension, plume dispersion, visual occlusion, and unstable artificial illumination, thereby creating substantial challenges for visual monitoring and perception in deep-sea mining systems [4].

Visual data for mining-system development are difficult to acquire: sea trials are costly, operational windows are limited, visibility is poor, and annotation is labor-intensive [5, 6]. Dense plumes, equipment occlusion, and abnormal interactions may be sparsely represented in recorded footage. Synthetic generation offers a way to construct controlled examples of these conditions [7].

These conditions motivate prompts that specify viewpoint, lighting, material interactions, and temporal evolution. Underwater datasets and physics-based simulators address complementary aspects of the problem [8, 9]; text-to-video (T2V) and image-to-video (I2V) generation offer an additional source of controlled visual examples.

We examine how prompt structure affects four T2V systems in five mining-relevant scenarios. The study combines an archived system-level comparison with a controlled-prompt extension and repeated requests, reports six VBench metrics, and evaluates domain-rule compliance in a system-masked subset. Its central question is whether prompt changes consistently raise the measured scores under the tested configurations. The resulting evidence supports first-stage selection and screening of candidate videos; operational transfer and plume-physics validation require different reference data and endpoints.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

2 Related Work

2.1 Deep-sea mining visual monitoring and underwater synthetic data

Underwater optical imaging is affected by absorption, scattering, wavelength-dependent attenuation, and backscatter, which can reduce contrast and detail [6, 10]. Image-enhancement and object-detection datasets, synthetic-image datasets, and simulators support underwater perception research [8, 11, 12, 13, 9, 14, 15]. Prompt-generated video adds controllable temporal scenarios to this landscape. Its visual screening, physical calibration, and contribution to perception training are distinct evaluation stages.

2.2 Video generative models and prompt engineering

T2V and I2V models offer a flexible means of synthesizing rare and controllable visual conditions without constructing complete physical simulation environments [16, 17, 18, 19, 20]. Video generation must nevertheless preserve spatial fidelity, temporal coherence, object consistency, motion continuity, and text–video alignment, all of which remain central challenges for diffusion-based video generation [21, 18]. These requirements are more stringent in engineering applications, where generated videos must satisfy physical and operational constraints in addition to visual-quality criteria. CogVideoX, Wan, Mochi-1, and closed-source commercial systems differ in accessibility, context length, deployment cost, inference interface, and prompt-processing behavior [19, 20, 22, 23]. Prompt engineering is therefore important for translating deep-sea mining constraints into generation instructions [24]. Long prompts can provide rich environmental and physical detail but may introduce redundancy; model-adapted prompts can emphasize core actions and state evolution but may omit useful contextual cues. Studies of long-context language models also suggest that outputs can be sensitive to the position of relevant information, motivating an analysis of prompt-position sensitivity in video generation [25, 26].

2.3 Evaluation of generated videos and research gap

Generated videos require spatial and temporal evaluation. VBench measures subject and background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality [27]; complementary benchmarks examine content alignment, spatial relationships, actions, and interactions [28, 29, 30, 31]. A clip can score well on general metrics yet depict sunlight, an external camera, or irrelevant objects in a deep-sea scene. We therefore report VBench scores and domain-rule judgments separately [32]. Shared prompts and repeated requests test metric sensitivity in the extension; independent ratings assess rule compliance only in an archived subset.

3 Materials and Methods

The experiments addressed three questions: how four configured T2V systems scored on general video-quality metrics; how shared, detailed, negative-constrained, and model-adapted prompts changed those scores within each system; and how often archived videos breached prespecified visual-monitoring rules. The scenarios covered camera viewpoint, illumination, sediment resuspension, plumes, and equipment–seabed interactions.

3.1 Experimental tools and environment

The archived 300-video comparison used two API systems and two locally deployed systems; complete historical service and checkpoint revisions were not recoverable. The 240-video extension has separate generation records for every clip. Jimeng used the `jimeng_t2v_v30` interface, Wan used `wanx2.1-t2v-plus`, Mochi used a local `mochi-1-preview` Diffusers directory, and CogVideoX used the `kijai/CogVideoX-5b-1.5-T2V` ComfyUI workflow. Interface names identify requests, not immutable weight revisions; the two local installations also lack captured immutable checkpoint revisions. Both local systems ran on NVIDIA H20 servers, and the new VBench evaluation ran on an RTX 4090 server.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

Table 1. Recorded settings for the controlled-prompt extension.

System	Access	Recorded request or inference settings	Unavailable detail
Jimeng Video 3.0	API	`jimeng_t2v_v30`; 121 requested frames; 16:9; seeds 20260901–20260903	Served weight revision, sampling steps, guidance, and server-side processing
Wan 2.1 T2V Plus	API	`wanx2.1-t2v-plus`; 1280×720; requested 5 s; prompt expansion off; seeds 20260901–20260903	Served weight revision, sampling steps, guidance, and server-side processing
Mochi 1 Preview	Local	×480; 80 frames at 16 fps; 50 steps; guidance 4.5; same three seed labels	Immutable checkpoint revision
CogVideoX 1.5 5B	Local	×768; 81 frames at 16 fps; 50 steps; guidance 6.0; CogVideoXDDIM; same three seed labels	Immutable checkpoint revision

An MP4-header audit of all 240 extension outputs found one specification per system: Jimeng, 1248×704 at 24 fps with 121 decoded frames; Wan, 1280×720 at 30 fps with 161 frames; Mochi, 848×480 at 16 fps with 79 frames; and CogVideoX, 1400×792 at 16 fps with 81 frames. Thus, Mochi decoded one frame fewer than the 80 requested, and CogVideoX's encoded resolution differed from the 1360×768 workflow setting. Per-file properties are supplied in `controlled_media_inventory_240.csv`; requested settings are not substituted for observed file specifications.

The saved local environment records identify Mochi's Diffusers version as 0.40.0 and the SHA-256 of its `model_index.json`; they identify the CogVideoX ComfyUI and wrapper commits as `ab885b33` and `8d6e53b5`, respectively. These identify the software environment and model configuration, but not immutable hashes of all weight tensors.

3.2 Experimental scenarios and variable design

The original prompt library indexed five scenarios, five style codes, and three numbered variants per scenario–style combination. The variant indices were not passed as distinct focus instructions in the located generation script. Let $\mathcal{S}$, $\mathcal{C}$, and $\mathcal{R}$ denote these three index sets. The library contained

$$|\mathcal{P}_{75}| = |\mathcal{S}|\,|\mathcal{C}|\,|\mathcal{R}| = 5 \times 5 \times 3 = 75. \qquad (1)$$

This indexing provided scenario coverage and prompt diversity. Across four systems, the baseline comprised

$$N_{\text{base}} = |\mathcal{M}|\,|\mathcal{P}| = 4 \times 75 = 300, \qquad (2)$$

where $\mathcal{M}$ denotes the evaluated systems. These are video outputs rather than unique prompt texts: the API systems shared a 75-record library, while the two local systems each used 75 records from a separate 225-record specialist library. The unused specialist records do not represent additional videos.

The archived Jimeng and Wan generation scripts point to `deepsea_prompts_75.json`; its texts match the saved VBench mappings. For Mochi and CogVideoX, the M- and X-prefixed records in `deepsea_prompts_specialized_225.json` match their mappings. Both files predate the corresponding videos. Original local-model generation commands were not recovered, so their exact generation-time inputs remain less certain than their evaluation-time mappings.

Five operating conditions were considered. S1, Hydrothermal Vent Approach, used a forward-facing main-camera perspective and emphasized extremely low visibility and turbulent hydrothermal plumes. S2, Soft-Clay Seabed Traversal, used a chassis-monitoring perspective and focused on crawler sinkage, shear failure, and sediment suspension [33]. S3, Polymetallic Nodule Collection, used a close-up view of the bucket suction head and represented interactions among the collection mechanism, sediment, and gravel. S4, Biological-Community Interference, used a partially obstructed camera perspective to represent occlusion caused by aggregating phototactic organisms. S5, Tailings-Discharge Monitoring, used a rear-facing monitoring perspective and focused on the dispersion of high-concentration mud plumes discharged from tailings pipes [3, 34, 35].

The style codes represent prompt-construction conventions coupled to the evaluated systems. Scenario descriptions and archived prompt texts, rather than inferred variant meanings, define the baseline inputs.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

The archived prompts were audited for internal inconsistencies. In the prespecified 60-video re-rating subset, one English prompt contains an unintended Chinese token, two positive prompts mention surface lighting, and 15 CogVideoX prompts specify a nominal 10-s timeline for clips approximately 6.1 s long. These inputs were retained and linked to their outputs so that the audit reflects the videos actually generated.

Representative prompt examples and word-order prompt variants are provided in the Electronic Supplementary Material.

3.3 Archived dual-track prompt implementation

3.3.1 Overview of prompt-generation strategies

An LLM-assisted workflow constructed scene descriptions containing camera, sediment, plume, and lighting constraints [24]. Archived scripts name `qwen-plus`, `doubao-seed-1-6-`, and `hunyuan-pro`; the saved texts define the evaluated inputs.

3.3.2 Perspective constraints and hallucination suppression

The prompts requested a restricted onboard camera field of view and excluded external views, complete vehicle bodies, sunlight, sky, the water surface, shallow-water vegetation, and land. Negative text also named watermarks and severe distortions. These were generation instructions, not guarantees of output compliance.

3.3.3 Video-generation models and dual-track adaptation

The archived experiment used long structured prompts for API-accessed Jimeng-V3 [23] and Wan-2.1 [20]. Jimeng-V3 is the reported service label; the new run records the provider's `jimeng_t2v_v30` 720P text-to-video request identifier, not a publicly fixed weight revision. It used model-adapted prompts for locally deployed Mochi-1 [22] and CogVideoX-1.5 [19]. Track assignment followed the original workflow and was not randomized.

For the API-accessed track, Jimeng-V3 used a rendering-oriented template comprising quality tags, camera parameters, lighting settings, and detailed scenario descriptions, whereas Wan-2.1 used a documentary-monitoring template emphasizing objective scene descriptions, texture, and noncinematic constraints. For the locally deployed track, Mochi-1 used a motion-consistency template that highlighted core actions, spatial relationships, and continuous motion, while CogVideoX-1.5 used a temporal-coherence template focused on state evolution and core background constraints.

Here, "model-adapted" denotes restructuring around motion and state evolution, not necessarily fewer words. Because prompt track and system identity were coupled, the archived comparison is a system-level benchmark.

The archived workflow comprised five phases; the controlled extension was conducted separately:

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

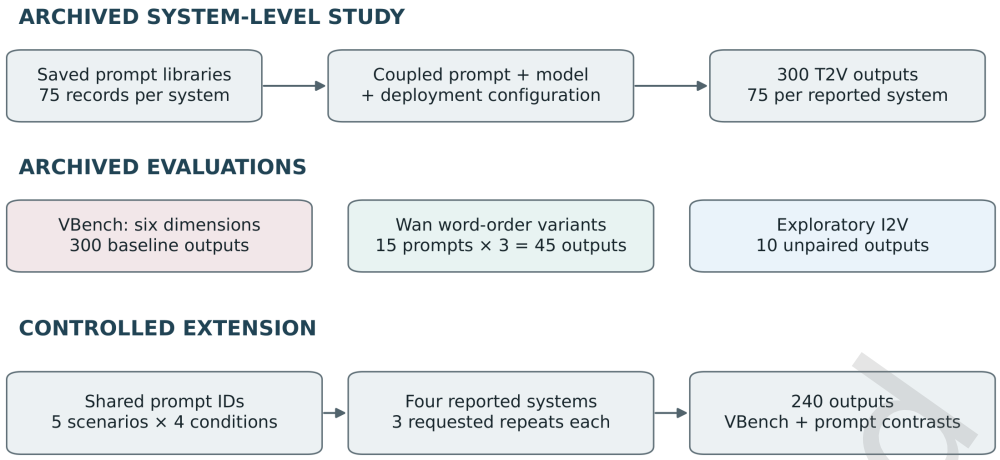


Figure 1. Study design. The archived baseline compares configured model–prompt–deployment systems; the controlled extension uses shared prompt IDs and repeated generation. The archived I2V outputs are exploratory and unpaired.

Phase 1—Prompt construction and dual-track adaptation. The archived API library contains 75 ID–text records; the two local libraries contain 75 each. Their saved evaluation mappings were checked. The new controlled library was frozen separately and linked to individual generation requests.

Phase 2—Generation: The API-library records were associated with Jimeng-V3 and Wan-2.1; the adapted libraries were associated with Mochi-1 and CogVideoX-1.5. The archive contains 75 baseline videos per system, or 300 total (Eq. (2)).

Phase 3—VBench evaluation: VBench was used to evaluate all 300 baseline videos. Six dimensions were recorded: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. Imaging-quality values were divided by 100 so that all reported dimensions were on a $[0,1]$ scale. Bar charts and box plots summarize system-level means and prompt-level score distributions [27, 29, 30].

Phase 4—Word-order experiment: Fifteen indexed prompts spanning five scenarios were selected for a Wan-only test. Three word-order variants were constructed for each prompt while retaining core semantic content, producing $N_{\text{order}} = 15 \times 3 = 45$ videos.

Variant 1 (Action-First) placed physical actions and events before environmental and lighting descriptions. Variant 2 (Environment-First) placed the deep-sea environment and lighting before the core actions. Variant 3 (Scattered-Integration) combined quality, camera, lighting, and action descriptions in continuous natural-language sentences. The videos were evaluated with the same VBench metrics, and between-variant differences were analyzed descriptively.

Phase 5—Exploratory I2V material: Ten archived I2V outputs used reference images and text. The surviving script names `wan2.6-i2v-flash`, but the execution-time API response is missing. Their content is not a verified homogeneous S1 set, and no paired T2V comparison is retained.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

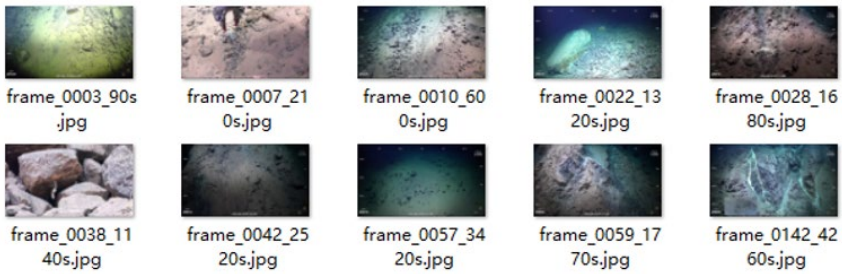


Figure 2. Archived reference frames used for exploratory image-to-video testing; their S1 scenario labels are not independently verified.



Figure 3. Representative thumbnails of image-to-video generated samples.

3.4 Data recording and processing

Each extension video has a generation sidecar recording its task ID, scenario, condition, repeat, requested seed, prompt text, model request, provider task ID when available, and SHA-256. Filename links the sidecar to six per-video VBench scores. The data schema is

$$\mathcal{D}_i = \{\text{task ID}, \text{scenario}, \text{condition}, \text{model}, \text{seed}, \\ \text{request}, \text{video SHA-256}, \text{six scores}\}.$$

The new records were grouped by system, prompt condition, and scenario for descriptive analysis. The historical baseline has less complete generation-side documentation. Archival tables and the 240-row machine-readable extension table accompany this revision.

3.5 Controlled-prompt and repeated-seed extension

The extension crossed four systems, five scenarios, four prompt conditions, and three requested seeds (20260901–20260903), yielding $4 \times 5 \times 4 \times 3 = 240$ videos. P1 was short; P2 used the same detailed positive text across systems; P3 added fixed exclusions to P2; and P4 combined model-adapted positive text with those exclusions. The numeric seed was a within-system repeat identifier, not a shared latent sample across architectures. Jimeng received exclusions appended to its prompt, whereas Wan and the local workflows used separate negative-conditioning fields. Thus, P3 is comparable as a requested constraint but not as an identical implementation across services.

Let x_{mcsrd} be the normalized VBench score for system m, condition c, scenario s, repeat r, and dimension d. We summarize each scenario–condition cell over its three generated videos and then average equally across the five scenarios:

$$\bar{x}_{mcsd} = \frac{1}{3}\sum_{r=1}^{3} x_{mcsrd}\,, \qquad \bar{x}_{mcd} = \frac{1}{5}\sum_{s=1}^{5} \bar{x}_{mcsd}\,. \tag{3}$$

For a within-system contrast, the descriptive paired mean difference is

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

$$\Delta_{m,d}^{(c-c')} = \frac{1}{15}\sum_{s=1}^{5}\sum_{r=1}^{3}(x_{mcsrd} - x_{mc'srd})\,. \qquad (4)$$

To separate repeat-to-repeat variation from scenario variation, we computed the sample standard deviation across three seeds in each fixed system–condition–scenario cell and averaged the 20 cell SDs for each system and dimension:

$$s_{mcsd} = \sqrt{\frac{1}{2}\sum_{r=1}^{3}(x_{mcsrd} - \bar{x}_{mcsd})^2}, \qquad \bar{s}_{md} = \frac{1}{20}\sum_{c=1}^{4}\sum_{s=1}^{5} s_{mcsd}\,. \qquad (5)$$

This is a descriptive measure of variation among three requested seeds for the tested cells, not uncertainty of a population-level mean. Dispersion across all 15 videos per condition also includes scenario variation. Contrasts are descriptive because the study has five selected scenarios and multiple outcomes.

The archived VBench checkout is based on commit 07bc8a4b74d5e0a23de42ed5880b899a1ff705f0; source-file hashes record local modifications. The cloud run used Python 3.12.3, PyTorch 2.7.1+cu118, and an RTX 4090. All 240 videos were evaluated in `custom_input` mode on six dimensions. Subject consistency, background consistency, aesthetic quality, and imaging quality passed all decoded frames to their metric implementations. Motion smoothness decoded all frames and used alternating frames for interpolation; dynamic degree used frame indices $0, k, 2k, \ldots$, where $k = \max\{1, \mathrm{round}(\mathrm{fps}/8)\}$. MUSIQ imaging quality used `longer` preprocessing (longest side at most 512 pixels) and was divided by 100 in the per-video table. The full-info JSON, filename-to-prompt maps, logs, and source manifest document the run. For one Mochi pilot clip evaluated on both local and cloud machines, the largest absolute score difference was 2.2×10^{-5} across six dimensions.

4 Evaluation Metrics and Experimental Protocol

4.1 VBench-based general video-quality metrics

VBench was used to measure six general video properties: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality [27]. Dynamic degree indicates detected motion, not physical correctness or an unconditionally desirable level of activity for monitoring. Let $x_{imd} \in [0,1]$ denote the score of video i, generated by model m, on dimension d. Raw imaging-quality values were divided by 100; the other dimensions were used on their reported [0,1] scales. For $n_m = 75$ baseline videos from model m, the sample mean and sample standard deviation were calculated as

$$\bar{x}_{md} = \frac{1}{n_m}\sum_{i=1}^{n_m} x_{imd}\,,$$

$$s_{md} = \sqrt{\frac{1}{n_m - 1}\sum_{i=1}^{n_m}(x_{imd} - \bar{x}_{md})^2}\,.$$

The model-level performance vector used in the dimension-wise summaries was

$$\boldsymbol{v}_m = [\bar{x}_{m1}, \bar{x}_{m2}, \ldots, \bar{x}_{m6}]^{\mathrm{T}}. \qquad (8)$$

The baseline standard deviations describe dispersion among the selected prompts and scenarios, not uncertainty from repeated generation at a fixed prompt. No weighted or unweighted composite quality index was used because the six dimensions represent distinct properties. Because VBench was not designed specifically for deep-sea mining, it formed the general-quality layer, while domain validity was assessed separately using the criteria in Section 4.2 [27, 29, 30, 31].

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

Table 2. VBench-based general video-quality metrics used in this study.

Metric	Evaluation focus in this study
Subject consistency	Whether the main subject maintains a stable appearance and structure across frames.
Background consistency	Whether the deep-sea background remains stable without non-physical jumps or severe flickering.
Motion smoothness	Whether object motion is continuous without obvious stuttering, frame loss, or unnatural teleportation.
Dynamic degree	Whether the video contains dynamic changes in sediment, plumes, lighting, or equipment motion.
Aesthetic quality	Overall visual perception, including lighting layout and color composition.
Imaging quality	Technical image quality, including clarity, noise, distortion, artifacts, and blur.

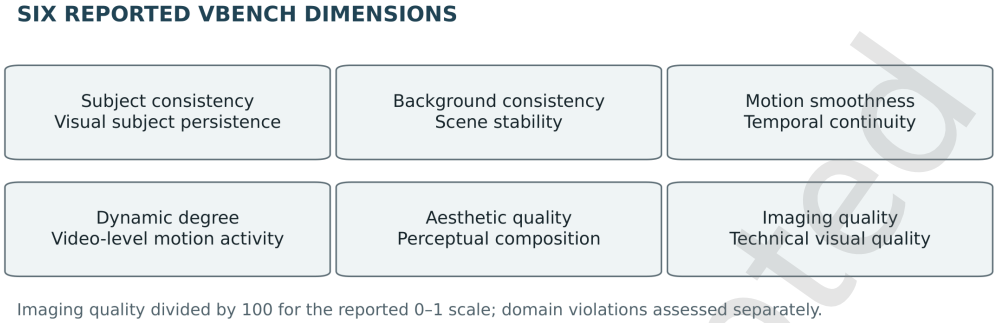


Figure 4. Six VBench dimensions reported in this study. Domain-rule violations were assessed separately and are not included in the six-dimensional score.

4.2 Rule-based domain violation analysis

Although VBench provides useful general-purpose video-quality metrics, it does not directly assess compliance with deep-sea mining constraints. A rule-based violation analysis was therefore used to identify obvious domain errors that could reduce the utility of synthetic videos for visual monitoring. The selected violations reflect requirements for first-person monitoring views, the absence of shallow-water elements, plausible visibility reduction by sediment plumes, and temporal continuity under artificial illumination [3, 34, 5, 36, 37, 38].

Four binary indicators were used: perspective violation, deep-sea environment violation, irrelevant-entity hallucination, and severe temporal artifact. For the new three-rater subset, let $z_{imk} \in \{0,1\}$ denote the majority label for video i from model m and violation type k, where 1 indicates that at least two raters marked a violation. The model-specific rate for violation k and the overall rate of videos with at least one majority-labeled violation were calculated as

$$R_{mk} = \frac{1}{n_m^{(v)}} \sum_{i=1}^{n_m^{(v)}} z_{imk},$$

$$R_m^{(\text{any})} = \frac{1}{n_m^{(v)}} \sum_{i=1}^{n_m^{(v)}} \mathbb{I}\left(\sum_{k=1}^{4} z_{imk} > 0 \right),$$

where $n_m^{(v)} = 15$ and $\mathbb{I}(\cdot)$ is the indicator function. Pooled rates were calculated over all 60 inspected videos.

The archived report summarizes one review of 60 videos, 15 per system, using complete clips and frames near 10%, 30%, 50%, 70%, and 90% of each clip. Its per-video rating sheet was not recovered, so the aggregate counts are retained as historical single-rater observations. Inter-rater agreement cannot be calculated from those counts. The 240 extension videos have not undergone domain-rule review.

For an independent reliability assessment, we prespecified a distinct, system-masked subset of 60 archived clips: 15 per system, covering numbered variants 1–3 in each of S1–S5. These variant numbers define sampling strata, not semantically matched focus levels across the different prompt libraries. Three independent raters received the same clips in separately randomized orders. They saw masked filenames, a standardized scenario card, and the

Just Accepted

positive instruction, but not system identity, VBench scores, or the negative-prompt field. Each rater recorded four binary indicators and a not-rateable flag after viewing the full clip; only playback failure qualified a clip as not rateable.

The prespecified analysis reports Fleiss' κ, pairwise Cohen's κ, mean pairwise raw agreement, and rater-specific positive rates for each indicator. Undefined κ values under constant marginal ratings would be marked not estimable. Consensus requires at least two of three positive ratings for each indicator; violation proportions use videos as the denominator and descriptive Wilson 95% intervals. Because the subset was selected by design, these intervals are not population-prevalence estimates. A video has an overall violation if any indicator reaches this majority threshold. The returned files contained 60 complete ratings per rater and no not-rateable entries.

Table 3. Rule-based domain violation indicators for synthetic deep-sea mining videos.

Indicator	Definition	Score
Perspective violation	Third-person view, complete vehicle body, external viewpoint, or drone-like view appears.	0/1
Deep-sea environment violation	Sunlight, blue sky, water surface, land, shallow-water vegetation, or other non-deep-sea elements appear.	0/1
Irrelevant entity hallucination	Unprompted divers, ships, humans, unrelated animals, or irrelevant objects appear.	0/1
Severe temporal artifact	Obvious flickering, abrupt object disappearance, structural collapse, or strong frame-to-frame inconsistency appears.	0/1

5 Results

5.1 Baseline dual-track performance

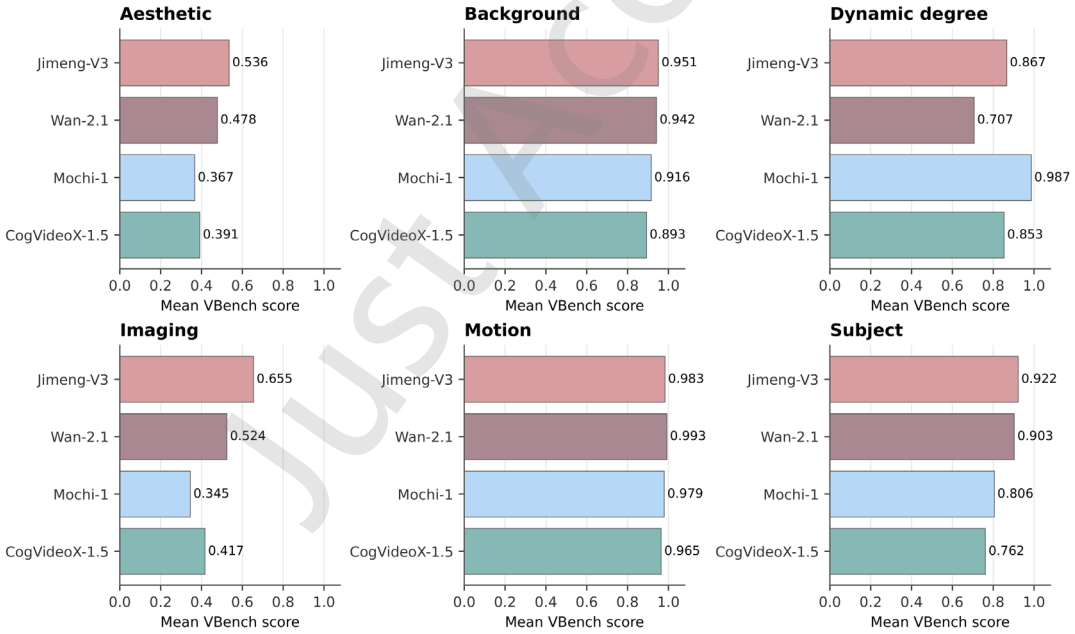


Figure 5. Archived baseline mean VBench scores across 75 outputs per reported system. Bars represent configured model–prompt–deployment systems, not isolated model effects; all plotted dimensions use the reported 0–1 scale.

Table 4. Archived baseline VBench scores by reported system label (mean ± standard deviation across 75 videos per system). Historical checkpoint identities are not independently verified.

Model	Aesthetic	Background	Dynamic	Imaging	Motion	Subject
Jimeng-V3	0.536 ± 0.059	0.951 ± 0.019	0.867 ± 0.342	0.655 ± 0.075	0.983 ± 0.010	0.922 ± 0.046

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

Model	Aesthetic	Background	Dynamic	Imaging	Motion	Subject
Wan-2.1	0.478 ± 0.070	0.942 ± 0.032	0.707 ± 0.458	0.524 ± 0.114	0.993 ± 0.005	0.903 ± 0.085
Mochi-1	0.367 ± 0.053	0.916 ± 0.025	0.987 ± 0.115	0.345 ± 0.087	0.979 ± 0.014	0.806 ± 0.064
CogVideoX-1.5	0.391 ± 0.052	0.893 ± 0.055	0.853 ± 0.356	0.417 ± 0.174	0.965 ± 0.033	0.762 ± 0.137

Figure 5 and Table 4 summarize the baseline VBench scores. In this system-level comparison, the API systems scored higher on aesthetic and imaging quality; the local systems showed higher dynamic-degree scores in some cases. Jimeng-V3 had the highest mean scores for aesthetic quality (0.536), imaging quality (0.655), background consistency (0.951), and subject consistency (0.922). Wan-2.1 had the highest motion-smoothness score (0.993), whereas Mochi-1 had the highest dynamic-degree score (0.987).

Within the API-accessed track, Jimeng-V3 scored higher than Wan-2.1 in five dimensions; Wan-2.1 scored higher in motion smoothness. Within the locally deployed track, CogVideoX-1.5 scored higher in aesthetic and imaging quality, while Mochi-1 scored higher in the other four dimensions. These baseline contrasts compare configured model–prompt–deployment systems.

5.2 Score distribution across prompt samples

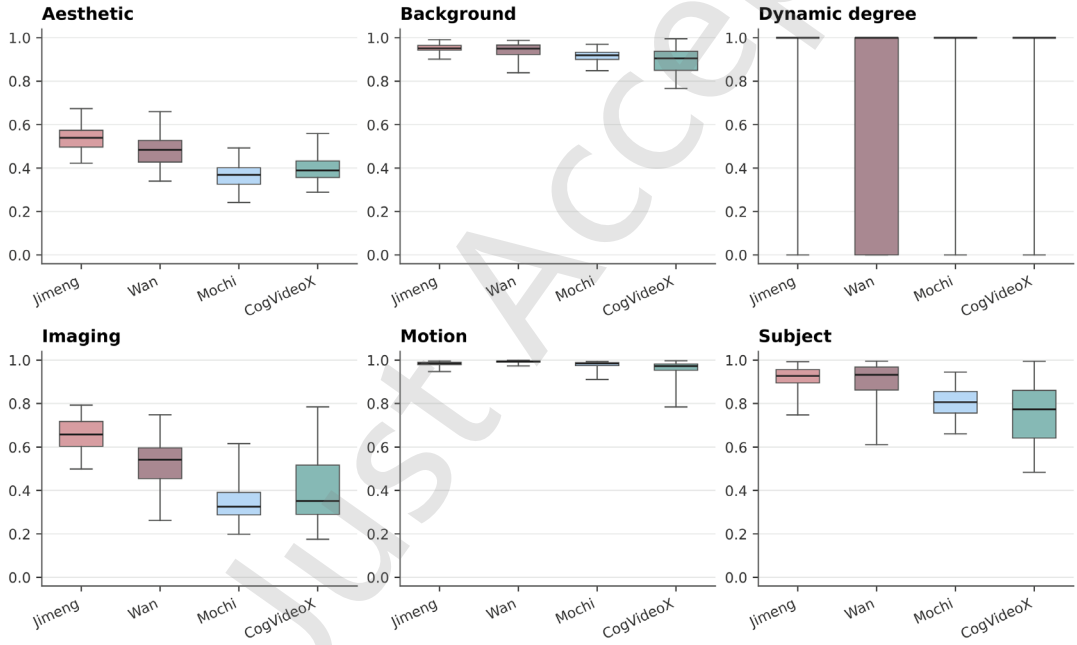


Figure 6. Prompt-level VBench score distributions (75 outputs per system). Boxes show the interquartile range and median; whiskers extend to the observed minimum and maximum. Dynamic degree is binary at the video level.

Figure 6 shows the score distributions across prompt samples. Jimeng-V3 and Wan-2.1 had relatively small standard deviations in background consistency (0.019 and 0.032, respectively) and motion smoothness (0.010 and 0.005), indicating limited prompt-level dispersion in these dimensions.

CogVideoX-1.5 showed the largest imaging-quality ($s = 0.174$) and subject-consistency ($s = 0.137$) dispersion. Dynamic degree varied most for Wan-2.1 ($s = 0.458$) and CogVideoX-1.5 ($s = 0.356$), partly reflecting the dimension's discrete behavior. These baseline standard deviations combine differences among prompts and scenarios; repeat-to-repeat variability is examined in the extension below.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

5.3 Controlled-prompt extension: four conditions and repeated seeds

All four systems yielded 60 scored files, with 15 per prompt condition. Each file was linked to a generation sidecar, video SHA-256, evaluation prompt, and six finite VBench scores. Table 5 reports condition means on a common 0–1 scale.

Table 5. Controlled-extension VBench means ($n = 15$ videos per system–condition; five scenarios and three seeds each). Column order follows the archived baseline table.

System	Condition	Aesthetic	Background	Dynamic	Imaging	Motion	Subject
Jimeng	P1	0.550	0.966	0.667	0.706	0.994	0.947
	P2	0.497	0.964	0.733	0.663	0.995	0.960
	P3	0.510	0.965	0.667	0.635	0.994	0.958
	P4	0.525	0.963	0.600	0.680	0.993	0.948
Wan	P1	0.523	0.969	0.733	0.544	0.995	0.956
	P2	0.481	0.959	0.600	0.529	0.996	0.953
	P3	0.481	0.959	0.600	0.529	0.996	0.953
	P4	0.445	0.962	0.600	0.459	0.996	0.957
Mochi	P1	0.373	0.957	0.133	0.315	0.995	0.930
	P2	0.399	0.958	0.133	0.404	0.995	0.958
	P3	0.384	0.953	0.000	0.360	0.992	0.945
	P4	0.398	0.956	0.333	0.361	0.994	0.947
CogVideoX	P1	0.416	0.961	0.400	0.519	0.985	0.944
	P2	0.385	0.963	0.200	0.523	0.995	0.950
	P3	0.401	0.965	0.467	0.455	0.996	0.960
	P4	0.380	0.962	0.600	0.441	0.996	0.949

Table 6. Mean within-cell sample SD across three requested seeds, averaged over 20 scenario–condition cells per system (Eq. (5)). Column order matches Table 5; these are not between-prompt SDs or confidence intervals.

System	Aesthetic	Background	Dynamic	Imaging	Motion	Subject
Jimeng	0.058	0.015	0.231	0.043	0.001	0.025
Wan	0.053	0.015	0.433	0.072	0.002	0.025
Mochi	0.032	0.011	0.202	0.079	0.001	0.018
CogVideoX	0.043	0.010	0.202	0.040	0.002	0.015

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

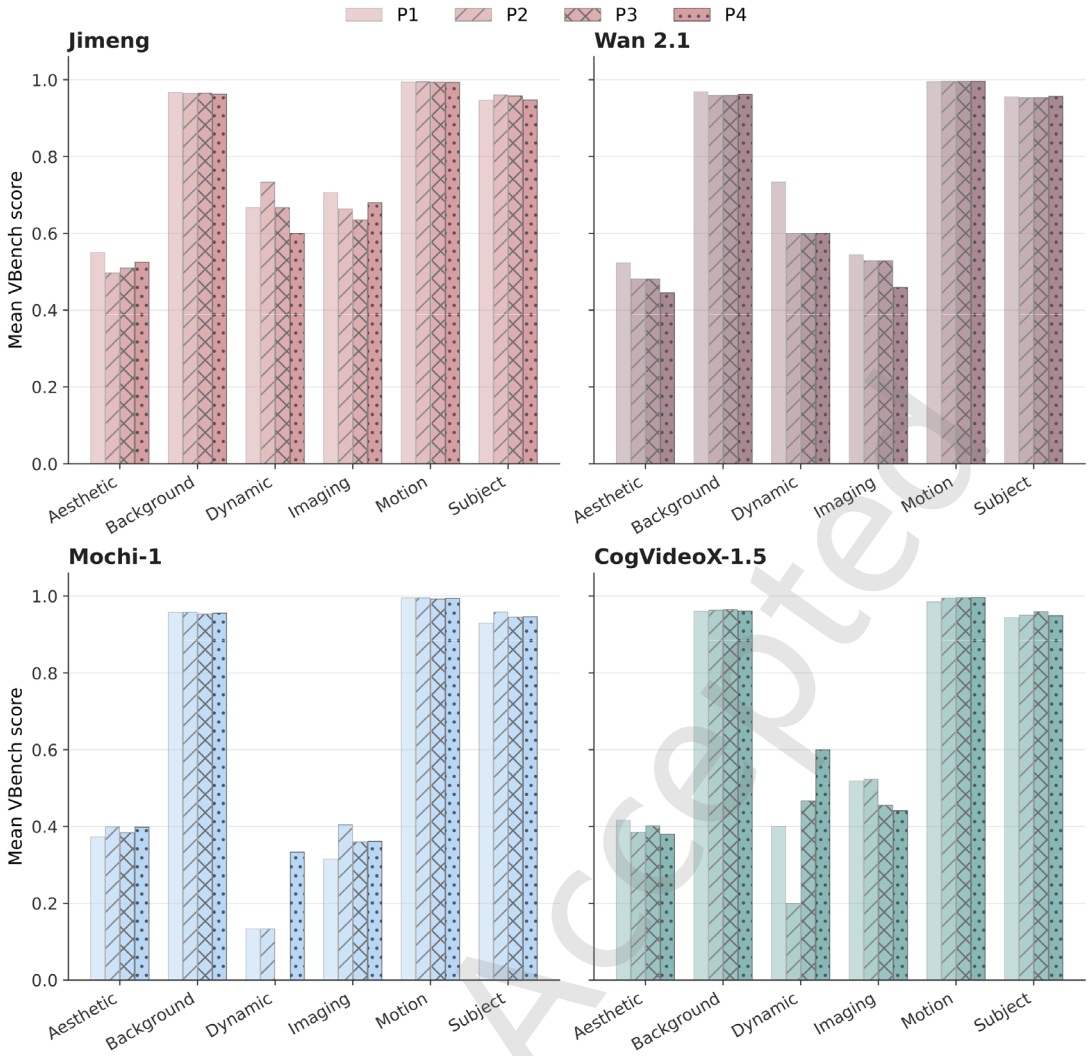


Figure 7. VBench means for the controlled-prompt extension, shown separately for each configured system. P1 is the short prompt, P2 the shared detailed positive prompt, P3 adds requested exclusions, and P4 uses model-adapted text with exclusions. Each bar summarizes 15 videos (five scenarios, three recorded seeds); hatching distinguishes conditions.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

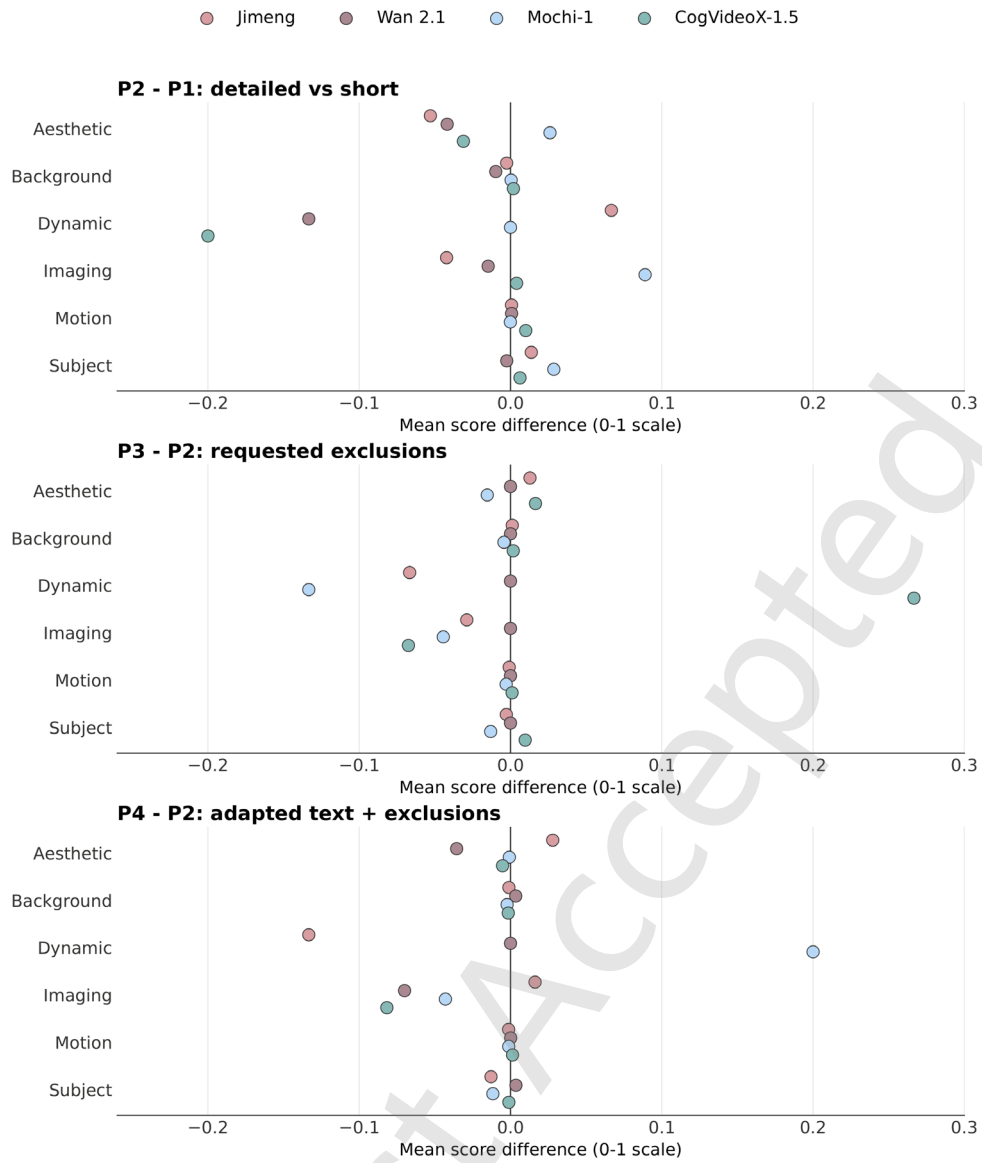


Figure 8. Descriptive within-system differences between prompt conditions across six VBench dimensions. Each point is a difference between two 15-video condition means, with scenario and repeat indices matched; positive values indicate a higher mean for the condition named first. The zero line denotes no mean difference.

Prompt detail had mixed effects (Figures 7–8). Relative to P1, P2 increased Jimeng's dynamic-degree mean by 0.067 but reduced its imaging-quality mean by 0.042. Mochi's imaging-quality mean rose by 0.089, whereas CogVideoX's dynamic-degree mean fell by 0.200. Table 6 reports observed repeat-to-repeat dispersion separately from variation among scenarios; dynamic degree is binary per clip and its SD should be interpreted accordingly. The shared P2 text fixed the positive instruction across systems; generation settings and service implementations still differed.

All 15 Wan P2/P3 scenario–seed pairs had identical scores and pixel-identical decoded frames, despite distinct task IDs, container hashes, and a nonempty negative-prompt field in the P3 requests. The service's handling of that field remains unknown. Elsewhere, P3–P2 dynamic-degree differences were -0.067 for Jimeng, -0.133 for Mochi, and $+0.267$ for CogVideoX (Figure 8). P4 also showed no consistent gain across the six dimensions. These results provide no general VBench advantage for longer or negatively constrained prompts; the extension has not yet been domain-rated.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

5.4 Domain-specific violation analysis

The independent review covered a prespecified, separately selected 60-video subset (15 per system). All three files contained one complete rating per system-masked video, with no not-rateable entries. Majority labels identified at least one violation in 37/60 clips (61.7%; descriptive Wilson 95% interval, 49.0–72.9%). The descriptive overall counts were 9/15 for Jimeng, 3/15 for Wan, 13/15 for Mochi, and 12/15 for CogVideoX; these small, system-specific strata were not used for inferential model ranking. Environment violations were the most frequent majority label (30/60). Table 7 reports the four criteria separately.

Table 7. Independent three-rater review of the distinct 60-video subset. Raw agreement is the mean of the three pairwise exact-agreement proportions; each majority count uses 60 videos as its denominator.

Violation criterion	Majority positive	Fleiss' κ	Raw agreement
Perspective	8 (13.3%)	0.338	81.1%
Deep-sea environment	30 (50.0%)	0.622	81.1%
Irrelevant entity	2 (3.3%)	0.117	83.3%
Severe temporal artifact	5 (8.3%)	0.000	72.2%

Agreement was criterion-dependent: Fleiss' κ was 0.622 for environmental violations but 0.117 for irrelevant entities and approximately zero for severe temporal artifacts. For the latter, one rater marked no positives, while the other two marked 13 and 17 of 60, respectively. The majority-based overall rate is therefore a transparent screening summary, not a high-certainty ground truth. The earlier single-rater aggregate (59/60) lacks recoverable per-video labels and cannot be matched clip by clip to the new subset; its original counts are retained in the Supplementary Material.

5.5 Exploratory word-order sensitivity analysis

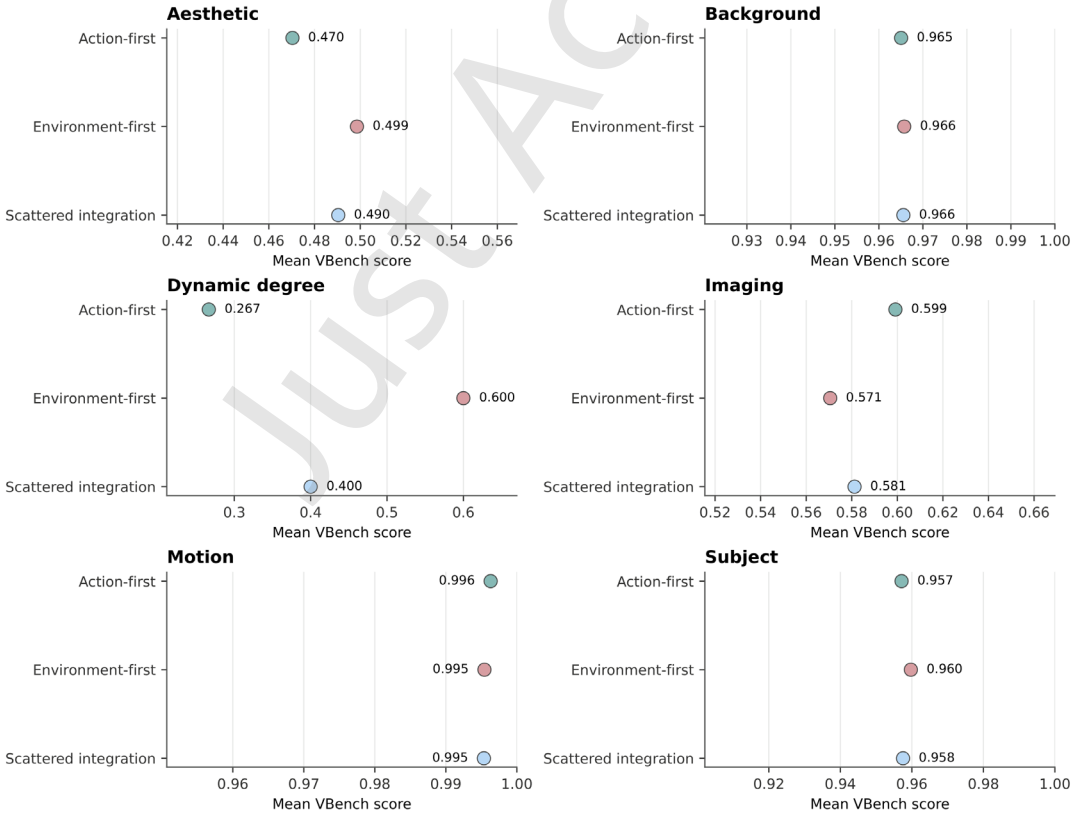


Figure 9. Archived Wan word-order experiment: mean VBench scores for each variant across 15 prompts. Axis ranges differ by panel to show the observed contrasts; compare values across panels using the printed scores rather than marker positions.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

The word-order results show an association between prompt variant and output scores; the variants altered information order but did not include repeated seeds. For variant v and dimension d, the descriptive contrast relative to the Action-First reference was defined as

$$\Delta_{vd} = \bar{x}_{vd} - \bar{x}_{\text{Action-First},d}. \qquad (11)$$

As shown in Figure 9, mean subject, background, and motion-smoothness scores differed by no more than 0.003 among variants. Across the five nondynamic dimensions, the largest absolute difference from Action-First was 0.029.

Dynamic degree showed the largest difference: Environment-First scored 0.600, Action-First 0.267, and Scattered-Integration 0.400. Thus, $\Delta_{\text{EF,dynamic}} = 0.333$ relative to Action-First in this 15-prompt sample. The result is an output-level association between information order and measured motion; this archived word-order experiment did not include seed repeats.

Figure 10 illustrates one WP2-S2 case: Environment-First emphasized clay deformation and sediment suspension under artificial light, whereas the other variants retained more vehicle-related content.

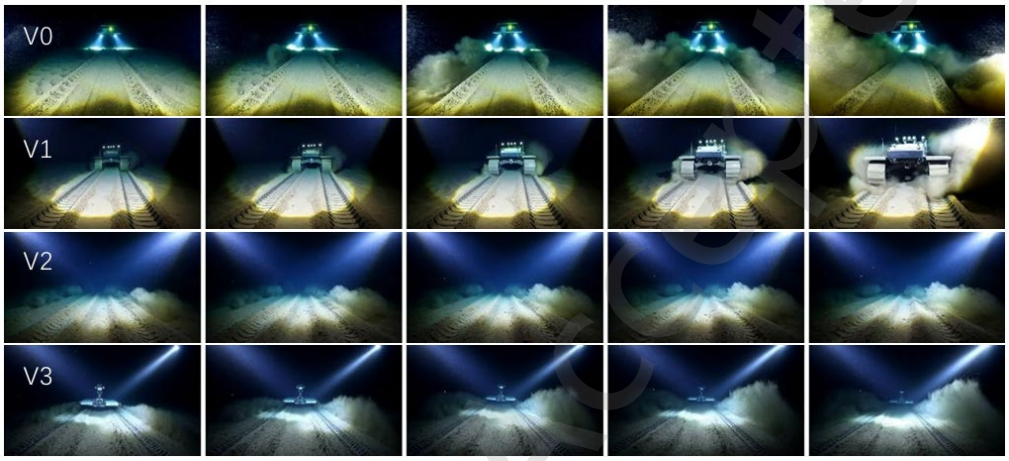


Figure 10. Video thumbnails comparing the WP2-S2 prompt variants.

Table 8. Prompt text variants for WP2-S2. V0 is the source prompt; only V1–V3 were generated and evaluated in the 45-video word-order experiment.

ID	Category	Main ordering strategy
V0	Original	Original modular order
V1	Variant 1	Mechanical action before environment
V2	Variant 2	Deep-sea environment and lighting before action
V3	Variant 3	Modifiers integrated into long natural sentences

5.6 Archived exploratory I2V material

The ten archived I2V clips remain descriptive score records, not a controlled T2V–I2V comparison. Their surviving request script names `wan2.6-i2v-flash`, whereas the original manuscript named Wan-2.1; the execution-time response was not recovered. Their images and prompts are not a verified homogeneous S1 set and are not paired with the 15 historical T2V clips. Accordingly, no between-mode effect size is reported.

6 Discussion

6.1 Implications for deep-sea mining visual monitoring

The controlled extension found prompt-dependent changes in general video-quality scores, but no prompt condition dominated across systems and dimensions. The independent review identified scene-level violations

and showed that agreement depends on the criterion. Together, these results support a staged use of generated clips: compare prompt conditions with general metrics, screen candidate clips against camera and domain rules, and reserve task-specific claims for tests on independently annotated real footage [39]. Physics-based simulators provide a complementary reference for plume behavior [9, 14, 15].

6.2 Conditional prompt-design guidance

The word-order sample linked environment-first wording to a higher dynamic-degree score, suggesting a focused follow-up rather than a universal template. The P1–P4 results favor choosing prompt structure for the target visual property and screening each clip against domain rules [24, 25, 26]. The pixel-identical Wan P2/P3 pairs raise a specific question about provider handling of negative requests; they show no negative-prompt effect for that service.

6.3 Scope and next validation stage

The archived comparison coupled model identity, prompt style, and deployment. Shared P2 wording improves comparability in the extension, although resolution, duration, inference settings, and API processing still differ. Three requested-seed outputs describe variation within tested cells. The archived single-rater aggregate cannot be paired with the new system-masked subset; agreement for irrelevant entities and severe temporal artifacts also indicates that those rule definitions need refinement before larger-scale labeling. The exploratory I2V files do not support a mode comparison.

The next stage would test two operational endpoints, neither measured in this study. For perception transfer, a larger real-video set should be annotated and split by independent dives or nonoverlapping segments; the same detector or tracker would then be trained on real footage alone and on real footage plus screened synthetic clips, with performance assessed only on held-out real footage [12, 7]. For plume behavior, visible plume-front displacement, expansion, and persistence could be compared over matched intervals against calibrated field sequences or numerical simulations; image-plane motion alone would not establish physical velocity without camera scale and flow information [3, 34, 5]. This staged design separates the present visual benchmark from physical and application-level validation.

7 Conclusion

Across an archived 300-video comparison and a 240-video controlled-prompt extension, within-system prompt contrasts varied by VBench dimension; no tested condition consistently raised all six scores. All 15 Wan P2/P3 pairs produced pixel-identical decoded frames despite distinct requests. A separate three-rater review identified at least one majority-labeled domain violation in 37/60 clips, with agreement varying substantially across criteria. The earlier single-rater aggregate (59/60) is not a directly comparable replicate.

These findings support metric-guided prompt testing and explicit domain screening for candidate deep-sea monitoring videos. The archived word-order and I2V analyses remain descriptive. The documented video corpus and evaluation protocol provide a basis for subsequent plume-kinematics and held-out real-video transfer tests.

Electronic Supplementary Material

Supplementary material containing prompt examples, and domain-rule definitions is available in the online version of this article at https://doi.org//10.26599/OCEAN.2026.9470033

Acknowledgements

None.

Author contribution statement

Nan Li is responsible for conceptualization, methodology, experiment implementation, data analysis, and writing of the original draft. Long Yu is responsible for supervision, methodology, manuscript review, and editing.

Just Accepted

Chaojun Huang, Yi Yang, and Kanglin Liu contributed to domain discussion, result interpretation, and manuscript revision. All the authors have approved the final manuscript.

Data availability

The 595 synthetic videos, prompt records, per-video VBench scores, available generation and evaluation provenance, de-identified human ratings, and analysis code are publicly available at https://zenodo.org/records/22936567.

Declaration of competing interest

All the contributing authors report no conflicts of interest in this work.

Funding

None.

Use of AI statement

During the preparation of this work, the authors used ChatGPT to assist with language polishing, translation, and structural revision suggestions. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the article.

References

[1] Kathryn A. Miller, Kirsten F. Thompson, Paul Johnston, and David Santillo. An overview of seabed mining including the current state of development, environmental impacts, and knowledge gaps. *Frontiers in Marine Science*, Volume 4 - 2017, 2018. ISSN 2296-7745. . URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2017.00418.

[2] Fei Sha, Mingshuai Xi, Xuguang Chen, Xuelin Liu, Hongying Niu, and Yuhang Zuo. A recent review on multi-physics coupling between deep-sea mining equipment and marine sediment. *Ocean Engineering*, 276: 114229, 2023. ISSN 0029-8018. . URL https://www.sciencedirect.com/science/article/pii/S0029801823006133.

[3] Fei Sha, Mingshuai Xi, Zhijie Wen, Xuguang Chen, Yuhang Zuo, Jingze Xu, Minglong Zhang, and Hongying Niu. A review on plumes generation and evolution mechanism during deep-sea polymetallic nodules mining. *Ocean Engineering*, 298: 117188, 2024. ISSN 0029-8018. . URL https://www.sciencedirect.com/science/article/pii/S0029801824005250.

[4] Wen Yao, Chi Tian, Yao Teng, Feng Diao, Xinguang Du, Peng Gu, and Weixin Zhou. Development of deep-sea mining and its environmental impacts: a review. *Frontiers in Marine Science*, Volume 12 - 2025, 2025. ISSN 2296-7745. . URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2025.1598584.

[5] Iason-Zois Gazis, Henko de Stigter, Jochen Mohrmann, Karl Heger, Melanie Diaz, Benjamin Gillard, Matthias Baeye, Mario E Veloso-Alarcón, Kaveh Purkiani, Matthias Haeckel, Annemiek Vink, Laurenz Thomsen, and Jens Greinert. Monitoring benthic plumes, sediment redeposition and seafloor imprints caused by deep-sea polymetallic nodule mining. *Nature Communications*, 16 (1): 1229, January 2025.

[6] Zhaorui Gu, Xiuhan Liu, Zhiqiang Hu, Guoyu Wang, Bing Zheng, John Watson, and Haiyong Zheng. Underwater computational imaging: a survey. *Intelligent Marine Technology and Systems*, 1 (1): 2, September 2023.

[7] Yaoming Zhuang, Longyu Ma, Jiaming Liu, Yonghao Xian, Baoquan Chen, Li Li, Chengdong Wu, Wei Cui, and Zhanlin Liu. A diffusion model-based image generation framework for underwater object detection. *Communications Engineering*, 5 (1): 22, December 2025.

[8] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *arXiv preprint arXiv:1901.05495*, January 2019.

Just Accepted

[9] Abdelhakim Amer, Olaya Álvarez-Tuñón, Halil Ibrahim Ugurlu, Jonas le Fevre Sejersen, Yury Brodskiy, and Erdal Kayacan. UNav-Sim: A visually realistic underwater robotics simulator and synthetic data-generation framework. *arXiv preprint arXiv:2310.11927*, October 2023.

[10] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters (RA-L)*, 5 (2): 3227–3234, 2020.

[12] Muwei Jian, Nan Yang, Chen Tao, Huixiang Zhi, and Hanjiang Luo. Underwater object detection and datasets: a survey. *Intelligent Marine Technology and Systems*, 2 (1): 9, March 2024.

[13] Reina Kaneko, Takumi Ueda, Hiroshi Higashi, and Yuichi Tanaka. PHISWID: Physics-inspired underwater image dataset synthesized from RGB-D images. *arXiv preprint arXiv:2404.03998*, April 2024.

[14] Musa Manhães, Sebastian Scherer, Martin Voss, Luiz Douat, and Thomas Rauschenbach. Uuv simulator: A gazebo-based package for underwater intervention and multi-robot simulation. In *OCEANS 2016 MTS/IEEE Monterey*, pages 1–8, 09 2016. .

[15] Easton Potokar, Spencer Ashford, Michael Kaess, and Joshua G. Mangelson. Holoocean: An underwater robotics simulator. In *2022 International Conference on Robotics and Automation (ICRA)*, page 3040–3046. IEEE Press, 2022. . URL https://doi.org/10.1109/ICRA46639.2022.9812353.

[16] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, October 2022.

[18] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

[20] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, March 2025.

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, April 2022.

[22] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

Just Accepted

[23] Volcano Engine. Jimeng AI Video Generation 3.0 720P: Text-to-Video API Documentation, 2025. URL https://www.volcengine.com/docs/85621/1792704?lang=zh. Official service documentation; accessed 24 September 2026.

[24] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, February 2024.

[25] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. . URL https://aclanthology.org/2024.tacl-1.9/.

[26] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.findings-naacl.130/.

[27] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[28] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, December 2018.

[29] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[30] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8406–8416, June 2025.

[31] Zelu Qi, Ping Shi, Shuqi Wang, Chaoyang Zhang, Fei Zhao, Zefeng Ying, Da Pan, Xi Yang, Zheqi He, and Teng Dai. T2veval: Benchmark dataset and objective evaluation method for t2v-generated videos. *Displays*, 91: 103178, 2026. ISSN 0141-9382. . URL https://www.sciencedirect.com/science/article/pii/S014193822500215X.

[32] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] Ning Zhang, Weikun Zhai, Shiyang Yin, Xuguang Chen, Jianxin Xia, Lubao Luan, and Aoshuang Zhai. Deep-sea soft soil and deep-sea mining vehicle interaction: From soil properties to locomotion performance. *Applied Ocean Research*, 154: 104372, 2025. ISSN 0141-1187. . URL https://www.sciencedirect.com/science/article/pii/S0141118724004930.

[34] Carlos Muñoz-Royo, Raphael Ouillon, Souha El Mousadik, Matthew H. Alford, and Thomas Peacock. An in situ study of abyssal turbidity-current sediment plumes generated by a deep seabed polymetallic nodule mining preprototype collector vehicle. *Science Advances*, 8 (38): eabn1219, 2022. . URL https://www.science.org/doi/abs/10.1126/sciadv.abn1219.

[35] Kun Huang, Yong Zhao, Weijia Xu, Han Lu, Tianbao Bai, and Li Zou. Research on numerical simulation of initial process of sediment discharge from deep-sea mining vehicle. *Ocean Engineering*, 327: 120973, 2025. ISSN 0029-8018. . URL https://www.sciencedirect.com/science/article/pii/S0029801825006869.

[36] Rudy Helmons, Lynyrd de Wit, Henko de Stigter, and Jeremy Spearman. Dispersion of benthic plumes in deep-sea mining: What lessons can be learned from dredging? *Frontiers in Earth Science*, Volume 10 - 2022, 2022. ISSN 2296-6463. . URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2022.868701.

Just Accepted

This is a Just Accepted version, keep updated with https://doi.org/10.26599/OCEAN.2026.9470033

OCEAN 2026, 2, 9470033

[37] Sabine Haalboom, Henko C. de Stigter, Christian Mohn, Thomas Vandorpe, Marck Smit, Laurens de Jonge, and Gert-Jan Reichart. Monitoring of a sediment plume produced by a deep-sea mining test in shallow water, málaga bight, alboran sea (southwestern mediterranean sea). *Marine Geology*, 456: 106971, 2023. ISSN 0025-3227. . URL https://www.sciencedirect.com/science/article/pii/S0025322722002420.

[38] Jeffrey C Drazen, Craig R Smith, Kristina M Gjerde, Steven H D Haddock, Glenn S Carter, C Anela Choy, Malcolm R Clark, Pierre Dutrieux, Erica Goetze, Chris Hauton, Mariko Hatta, J Anthony Koslow, Astrid B Leitner, Aude Pacini, Jessica N Perelman, Thomas Peacock, Tracey T Sutton, Les Watling, and Hiroyuki Yamamoto. Opinion: Midwater ecosystems must be considered when evaluating environmental risks of deep-sea mining. *Proc. Natl. Acad. Sci. U. S. A.*, 117 (30): 17455–17460, July 2020.

[39] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in machine vision. *Mach. Intell. Res.*, 21 (5): 831–869, October 2024.

Just Accepted