SeaCache: Spectral-Evolution-Aware Cache for Accelerating Diffusion Models

Supplementary Material

This document presents supplementary materials that could not be included in the main manuscript due to page limitations. We first present the deviation analysis of the linear diffusion process. We then provide additional experiments that further validate the effectiveness of the proposed SeaCache. Finally, we include a zipped archive that contains our implementation for *FLUX.1-dev*, *HunyuanVideo*, and *Wan2.1 1.3B* in the `code` directory, as well as additional video samples in the `video_samples` directory for qualitative comparison between SeaCache and baseline methods.

1. Derivation of Optimal Linear Response

To design a filter that reflects spectral evolution, we formalize how the effective frequency band changes across timesteps. Motivated by Spectral Diffusion [16] and Wiener Filtering [15], we adopt the timestep-dependent frequency response derived from the optimal linear denoiser $h_t^\star$. [1]

Setup and assumptions. We consider the linear mixture of iterative denoising generative models (DPMs and RFs) at timestep t,

$$x_t = a_t\, x_0 + b_t\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where x_0 is the clean signal, assumed to be wide-sense stationary, and ϵ is zero-mean white Gaussian noise with flat power spectral density $S_\varepsilon(f) = 1$. We also assume that x_0 and ϵ are independent.

The Fourier-domain version of Eq. (1) is

$$\mathcal{X}_t(f) = a_t\, \mathcal{X}_0(f) + b_t\, \mathcal{E}(f), \tag{2}$$

where $\mathcal{X}_0(f)$, $\mathcal{X}_t(f)$, and $\mathcal{E}(f)$ are the Fourier transforms of x_0, x_t and ϵ at frequency f, respectively.

The filter h_t estimates x_0 from x_t as

$$\widehat{x}_0 = h_t * x_t \quad \Longleftrightarrow \quad \widehat{\mathcal{X}}_0(f) = \mathcal{H}_t(f)\, \mathcal{X}_t(f), \tag{3}$$

where h_t is a linear reconstruction estimator, $\mathcal{H}_t(f)$ is the frequency response of h_t, and $\widehat{x}_0$, $\widehat{\mathcal{X}}_0(f)$ are the estimated signal and its Fourier counterpart, respectively.

We define the signal reconstruction MSE objective, which is equivalent to the denoising objective of diffusion models,

$$J_t = \left\| h_t * x_t - x_0 \right\|_2^2, \qquad h_t^\star = \arg\min_{h_t}\ \mathbb{E}[\, J_t\,], \tag{4}$$

where the expectation is taken over (x_0, ϵ).

[1] Throughout this section, the symbol "$*$" denotes convolution in the spatial (or spatio-temporal) domain, while the superscript "*" denotes complex conjugation of Fourier coefficients.

Frequency-domain MSE expansion. By Parseval's theorem [10], the reconstruction MSE (Eq. (4)) decomposes as an integral over frequencies. Since $\mathcal{H}_t(f)$ acts independently at each frequency, minimizing the total MSE is equivalent to minimizing $J_t(f)$ for every f,

$$J_t(f) = \mathbb{E}\Big[\ \big|\ \mathcal{H}_t(f)\, \mathcal{X}_t(f) - \mathcal{X}_0(f)\big|^2\ \Big]. \tag{5}$$

We now expand $J_t(f)$ using standard complex-valued quadratic expansion:

$$\begin{aligned} J_t(f) &= \mathbb{E}\Big[\ \big|\ \mathcal{H}_t(f)\, \mathcal{X}_t(f) - \mathcal{X}_0(f)\big|^2\ \Big] \\ &= \mathbb{E}\big[\ \big(\mathcal{H}_t(f)\, \mathcal{X}_t(f) - \mathcal{X}_0(f)\big)\big(\mathcal{H}_t(f)\, \mathcal{X}_t(f) - \mathcal{X}_0(f)\big)^*\ \big] \\ &= |\mathcal{H}_t(f)|^2\, \mathbb{E}\big[\ |\mathcal{X}_t(f)|^2\ \big] - \mathcal{H}_t(f)\, \mathbb{E}[\ \mathcal{X}_t(f)\, \mathcal{X}_0(f)^*\ ] \\ &\quad - \mathcal{H}_t(f)^*\, \mathbb{E}[\ \mathcal{X}_0(f)\, \mathcal{X}_t(f)^*\ ] + \mathbb{E}\big[\ |\mathcal{X}_0(f)|^2\ \big], \end{aligned} \tag{6}$$

where all quantities are evaluated at frequency f.

We next simplify the two expectation terms $\mathbb{E}[\mathcal{X}_0(f)\, \mathcal{X}_t(f)^*]$ and $\mathbb{E}\big[|\mathcal{X}_t(f)|^2\big]$, which will be used in the subsequent derivation. Let $S_x(f)$ denote the power spectrum of x_0. The first term can be written as

$$\begin{aligned} \mathbb{E}[\mathcal{X}_0(f)\, \mathcal{X}_t(f)^*] &= \mathbb{E}\big[\mathcal{X}_0(f)\, \big(a_t \mathcal{X}_0(f) + b_t \mathcal{E}(f)\big)^*\big] \\ &= a_t\, \mathbb{E}\big[|\mathcal{X}_0(f)|^2\big] + b_t\, \mathbb{E}[\mathcal{X}_0(f)\, \mathcal{E}(f)^*] \\ &= a_t\, \mathbb{E}\big[|\mathcal{X}_0(f)|^2\big] \\ &= a_t\, S_x(f), \end{aligned} \tag{7}$$

since we assume that x_0 is wide-sense stationary and independent of the noise ϵ, so $\mathbb{E}[\mathcal{X}_0(f)\, \mathcal{E}(f)^*] = 0$ in Eq. (7). Next, we expand the second expectation term:

$$\begin{aligned} \mathbb{E}\big[\ |\mathcal{X}_t(f)|^2\ \big] &= \mathbb{E}\big[\big(a_t \mathcal{X}_0(f) + b_t \mathcal{E}(f)\big)\big(a_t \mathcal{X}_0(f) + b_t \mathcal{E}(f)\big)^*\big] \\ &= a_t^2\, \mathbb{E}\big[|\mathcal{X}_0(f)|^2\big] + b_t^2\, \mathbb{E}\big[|\mathcal{E}(f)|^2\big] \\ &\quad + a_t b_t\, \mathbb{E}[\mathcal{X}_0(f)\, \mathcal{E}(f)^*] + a_t b_t\, \mathbb{E}[\mathcal{X}_0(f)^*\, \mathcal{E}(f)] \\ &= a_t^2\, \mathbb{E}\big[|\mathcal{X}_0(f)|^2\big] + b_t^2\, \mathbb{E}\big[|\mathcal{E}(f)|^2\big] \\ &= a_t^2\, S_x(f) + b_t^2\, S_\varepsilon(f) \\ &= a_t^2\, S_x(f) + b_t^2, \end{aligned} \tag{8}$$

since in Eq. (8), the cross terms vanish because of independence, $\mathbb{E}[\mathcal{X}_0(f)\mathcal{E}(f)^*] = \mathbb{E}[\mathcal{X}_0(f)^*\mathcal{E}(f)] = 0$, and whiteness of the noise ϵ implies $\mathbb{E}[|\mathcal{E}(f)|^2] = S_\varepsilon(f) = 1$.

Optimality by differentiation. Differentiating Eq. (6) with respect to $\mathcal{H}_t(f)^*$ using Wirtinger derivative [1] and setting the result to zero to find the optimal linear filter under the linear MMSE criterion, we obtain

$$\frac{\partial J_t(f)}{\partial \mathcal{H}_t(f)^*} = \mathcal{H}_t(f)\, \mathbb{E}\big[\ |\mathcal{X}_t(f)|^2\ \big] - \mathbb{E}[\mathcal{X}_0(f)\, \mathcal{X}_t(f)^*] = 0. \tag{9}$$

Table 1. Runtime overhead of SEA filtering per sample, averaged over 10 runs.

Model	SEA Filtering (s)	Latency (s)	Overhead (%)
FLUX (2D FFT)	0.058	9.4	0.6
HunyuanVideo (3D FFT)	0.362	90.8	0.4

Using Eqs. (7) and (8), the unique minimizer is

$$\mathcal{H}_t^\star(f) = \frac{\mathbb{E}[\mathcal{X}_0(f)\,\mathcal{X}_t(f)^*]}{\mathbb{E}[\,|\mathcal{X}_t(f)|^2\,]} = \frac{a_t\,S_x(f)}{a_t^2\,S_x(f) + b_t^2}, \tag{10}$$

where $\mathcal{H}_t^\star(f)$ is the Fourier transform of h_t^*. We define the optimal frequency response

$$G_t(f) \;\triangleq\; \mathcal{H}_t^\star(f). \tag{11}$$

Power-law prior. We adopt an empirical natural-image power-law assumption for the power spectrum [2, 3, 12, 13],

$$S_x(f) \simeq A\,|f|^{-\beta}, \tag{12}$$

where $A > 0$ is an amplitude scaling factor and β is a frequency exponent. In our experiments, we set $A = 1$ and $\beta = 2$ for images and $\beta = 3$ for videos. Substituting this prior into the optimal response in Eq. (10) gives

$$G_t(f) = \frac{a_t\,|f|^{-\beta}}{a_t^2\,|f|^{-\beta} + b_t^2}, \tag{13}$$

which shows that the effective passband widens as a_t increases (*spectral evolution*). Note that the SEA filter used in our method $G_t^{\text{norm}}(f)$ is a normalized variant of $G_t(f)$. Its form is provided in the main manuscript.

2. Runtime Overhead of SEA Filtering

At every sampling step, SeaCache inserts an additional FFT $\rightarrow$ frequency-domain filtering $\rightarrow$ iFFT pass to construct SEA-filtered features. Thus, we measure how much of the end-to-end sampling time this pass occupies under a 50% caching ratio, keeping all other settings identical to the main experiments. For *FLUX* [7, 8] with SeaCache, the SEA filtering pass takes on average 0.058 s per sample out of a total latency of 9.4 s, corresponding to only about 0.6% of the overall generation time. For HunyuanVideo [6] with SeaCache, the 3D FFT-based SEA filtering costs 0.362 s per sample while the total latency is 90.8 s, roughly 0.4% of the end-to-end runtime. As summarized in Table 1, the SEA filtering introduces a negligible runtime overhead while enabling substantially better preservation of the original outputs compared to prior caching schemes.

3. Additional Evaluation

3.1. Quantitative Comparison in T2V Generation

VBench on *HunyuanVideo*. We evaluate SeaCache against TeaCache [9] and TaylorSeer [4] on all VBench [5] dimensions (Tab. 2), where the upper rows correspond to the 50% refresh-ratio budget and the lower rows to the 30% budget. All detailed settings follow the main manuscript. Aggregating by average rank across dimensions (Tab. 4), SeaCache ranks first under both budgets, scoring 1.91 vs. $2.03/2.06$ at $\approx 50\%$, and 1.75 vs. $2.16/2.09$ at $\approx 30\%$. This indicates the strongest overall performance across VBench dimensions on *HunyuanVideo* .

VBench on *Wan2.1 1.3B*. We repeat the evaluation on all VBench dimensions for *Wan2.1* [14] (Tab. 3) with the same two budgets and the same experimental details as in the main manuscript. In aggregate (Tab. 5), SeaCache delivers stable performance across dimensions, ranking second under both budgets, 1.97 at $\approx 50\%$ (vs. the best 1.91) and 2.13 at $\approx 30\%$ (vs. the best 1.53). Although our cache configurations are designed to closely track the original full-refresh sampling trajectory, the VBench results on *Wan2.1* still show that SeaCache provides robust performance across dimensions and refresh-ratio budgets.

CompressedVQA on T2V. To further quantify how caching affects video quality, we report scores from CompressedVQA [11], a full-reference video quality assessment (VQA) metric. For each video, we treat the uncached trajectory as the reference and compute single-scale and multi-scale scores between the cached outputs. Tab. 6 summarizes the results on *HunyuanVideo* and *Wan2.1 1.3B* at two cache budgets with refresh ratios of approximately 50% and 30%. Across both models and budgets, SeaCache consistently achieves the highest single-scale and multi-scale scores among all caching baselines, indicating that it best preserves the visual quality of the original trajectory while still enjoying substantial reductions in the refresh ratio.

3.2. Qualitative Comparison in T2I Generation

In Fig. 1, we provide additional qualitative comparisons on *FLUX* at refresh ratios of approximately 50% (top panel) and 30% (middle panel), along with an additional set of examples at both cache budgets in the bottom panel.

At 50% refresh ratio in top-left of Fig 1, SeaCache preserves a clean water surface without the blocky artifacts or texture distortions that appear in the baselines. In the top-right example, the baselines either generate a blurry lemon or fail to capture the fluid dynamics inside the bottle, whereas SeaCache correctly synthesizes both the glass bottle and the orange liquid, closely matching the full-compute original reference.

At a more aggressive 30% refresh ratio in the middle panel, SeaCache again stays closest to the full-compute ref-

Table 2. VBench metrics in *HunyuanVideo*.

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
TeaCache (δ=0.12)	95.59%	95.99%	99.14%	98.77%	62.50%	60.92%	62.07%	**86.31%**
TaylorSeer ($\mathcal{S}$=2)	95.75%	96.20%	99.09%	98.83%	**63.89%**	**60.93%**	**62.73%**	83.47%
SeaCache (δ=0.19)	**95.77%**	**96.28%**	**99.15%**	**98.88%**	62.50%	60.55%	62.01%	85.28%
TeaCache (δ=0.2)	95.57%	96.04%	99.18%	98.76%	62.50%	60.28%	60.28%	**86.47%**
TaylorSeer ($\mathcal{S}$=3)	95.67%	96.18%	99.07%	98.86%	**63.89%**	**60.64%**	**63.25%**	82.20%
SeaCache (δ=0.35)	**95.78%**	**96.35%**	**99.20%**	**98.92%**	61.11%	60.00%	61.02%	82.59%

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Temporal Style	Appearance Style	Overall Consistency
TeaCache (δ=0.12)	**64.71%**	**96.00%**	89.61%	61.84%	**42.81%**	24.39%	19.85%	**26.91%**
TaylorSeer ($\mathcal{S}$=2)	58.38%	95.00%	**90.87%**	60.80%	40.48%	24.44%	**19.89%**	26.60%
SeaCache (δ=0.19)	63.64%	94.00%	90.26%	**62.96%**	40.92%	**24.66%**	19.83%	26.63%
TeaCache (δ=0.2)	**63.34%**	92.00%	89.81%	59.65%	**44.48%**	24.26%	19.93%	**26.68%**
TaylorSeer ($\mathcal{S}$=3)	60.06%	92.00%	89.26%	57.78%	41.72%	**24.35%**	20.02%	26.57%
SeaCache (δ=0.35)	58.38%	**94.00%**	**92.24%**	**60.63%**	42.88%	24.34%	**20.10%**	26.33%

Table 3. VBench metrics in *Wan2.1 1.3B*.

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
TeaCache (δ=0.09)	95.89%	**97.09%**	98.30%	97.37%	81.94%	**62.48%**	67.88%	80.46%
TaylorSeer ($\mathcal{S}$=2)	95.78%	96.90%	**98.37%**	**97.47%**	**88.89%**	62.14%	**68.08%**	**82.75%**
SeaCache (δ=0.2)	**95.96%**	97.05%	98.20%	97.41%	84.72%	62.31%	68.01%	81.17%
TeaCache (δ=0.15)	**96.04%**	**97.02%**	**98.21%**	97.35%	83.33%	**62.25%**	67.47%	**80.22%**
TaylorSeer ($\mathcal{S}$=3)	95.32%	96.54%	**98.21%**	**97.48%**	**84.72%**	60.85%	**67.83%**	78.32%
SeaCache (δ=0.35)	96.03%	97.00%	98.12%	97.39%	81.94%	61.71%	67.66%	79.75%

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Temporal Style	Appearance Style	Overall Consistency
TeaCache (δ=0.09)	52.67%	**72.00%**	92.95%	71.46%	23.91%	23.07%	20.06%	23.42%
TaylorSeer ($\mathcal{S}$=2)	53.73%	70.00%	91.22%	**75.48%**	**30.09%**	22.75%	**20.13%**	23.41%
SeaCache (δ=0.2)	**53.89%**	70.00%	**93.01%**	69.50%	22.89%	**23.32%**	20.04%	**23.51%**
TeaCache (δ=0.15)	51.91%	**72.00%**	**90.56%**	67.67%	**24.27%**	**22.98%**	20.09%	**23.58%**
TaylorSeer ($\mathcal{S}$=3)	45.05%	69.00%	87.83%	60.79%	20.20%	22.37%	**20.64%**	23.17%
SeaCache (δ=0.35)	**53.20%**	68.00%	89.67%	**69.57%**	23.62%	22.96%	20.06%	23.18%

Table 4. Comparison of avg. rank on VBench in *HunyuanVideo*.

Method ($\approx 50\%$)	Rank $\downarrow$	Method ($\approx 30\%$)	Rank $\downarrow$
TeaCache (δ=0.12)	2.03	TeaCache (δ=0.20)	2.16
TaylorSeer ($\mathcal{S}$=2)	2.06	TaylorSeer ($\mathcal{S}$=3)	2.09
SeaCache (δ=0.19)	**1.91**	SeaCache (δ=0.35)	**1.75**

Table 5. Comparison of avg. rank on VBench in *Wan2.1 1.3B*.

Method ($\approx 50\%$)	Rank $\downarrow$	Method ($\approx 30\%$)	Rank $\downarrow$
TeaCache (δ=0.09)	2.13	TeaCache (δ=0.15)	**1.53**
TaylorSeer ($\mathcal{S}$=2)	**1.91**	TaylorSeer ($\mathcal{S}$=3)	2.34
SeaCache (δ=0.30)	1.97	SeaCache (δ=0.35)	2.13

Table 6. CompressedVQA [11] scores on *HunyuanVideo* and *Wan2.1 1.3B* under single-scale and multi-scale settings.

HunyuanVideo			*Wan2.1 1.3B*		
Method ($\approx 50\%$)	Single-scale score ↑	Multi-scale score ↑	Method ($\approx 50\%$)	Single-scale score ↑	Multi-scale score ↑
TeaCache (δ=0.12)	2.72	2.76	TeaCache (δ=0.09)	2.97	3.03
TaylorSeer ($\mathcal{S}$=2)	2.92	2.95	TaylorSeer ($\mathcal{S}$=2)	1.90	1.95
SeaCache (δ=0.19)	**3.98**	**3.99**	SeaCache (δ=0.30)	**3.93**	**3.95**
Method ($\approx 30\%$)	Single-scale score ↑	Multi-scale score ↑	Method ($\approx 30\%$)	Single-scale score ↑	Multi-scale score ↑
TeaCache (δ=0.20)	2.11	2.16	TeaCache (δ=0.15)	2.44	2.49
TaylorSeer ($\mathcal{S}$=3)	2.22	2.26	TaylorSeer ($\mathcal{S}$=3)	1.38	1.42
SeaCache (δ=0.35)	**3.13**	**3.17**	SeaCache (δ=0.35)	**3.09**	**3.11**

erence. In the middle-left example, only SeaCache reconstructs seven well-formed stars consistent with the original, while competing methods either miss or severely deform several stars. In the middle-right example, SeaCache produces five chopsticks with consistent length and color, whereas the baselines generate chopsticks with mismatched geometry and appearance.

In the bottom panel of Fig. 1, we further compare the same text prompts across different cache budgets using the same seed. In the top row of the panel, for the prompt requesting exactly the word "CUBE," the baselines repeatedly hallucinate cube-like patterns in the background, whereas SeaCache is the only method that successfully renders the intended text. In the last row of the panel, all methods generate six wooden ice creams, but the baselines produce slightly different designs or colors compared to the full-compute reference, while SeaCache most closely matches the original design.

These additional cases further support that SeaCache best preserves the original content and layout while operating under the same cache budgets.

3.3. Qualitative Comparison in T2V Generation

Fig. 2 present further qualitative comparisons on *HunyuanVideo* and *Wan2.1 1.3B*, respectively. For each prompt, we horizontally concatenate the same intermediate frame index from the full-compute reference and all caching variants to isolate per-frame differences. On *HunyuanVideo* at a 30% refresh ratio, the baselines exhibit severe artifacts around the hands during the Taichi motion, while SeaCache preserves a plausible pose with smooth limb contours. At 50% refresh, the baselines render a skateboard that appears to float above the surfboard, whereas SeaCache correctly places the skateboard in contact with the surfboard, matching the original video, as shown in the right side of Fig. 2.

On *Wan2.1 1.3B* at a 30% refresh ratio the baselines introduce noticeable distortions near the truck wheels and bicycles, but these artifacts do not appear in the SeaCache outputs, as visualized in Fig 2. At 50% refresh, competing methods either cause food items on the table to disappear or introduce artifacts on the panda, while SeaCache closely follows the full-compute trajectory without these failures. Overall, these qualitative results indicate that SeaCache better tracks the original dynamics and adheres more faithfully to the text prompts while avoiding objectionable artifacts.

Note that all videos corresponding to Fig. 2 are included in the `video_samples` directory of the supplementary material.

4. Limitation

To derive the optimal linear filter, we adopt several simplifying assumptions that make the spectral response analytically tractable, even though they need not hold exactly in practice. We model the signal spectrum with a power law under a radial view, whereas generated samples, particularly at later timesteps or in highly synthetic backgrounds with no salient objects, can deviate from this behavior. We also assume wide-sense stationarity and independence between signal and noise. When these conditions are violated, the closed-form linear filter is no longer strictly optimal and can introduce bias.

In addition, our analysis is formulated in the image or video domain, while most modern generative models operate in a learned latent space. The encoder can reshape the spectrum, so the latent distribution may differ from the assumed pixel-domain power-law model, and our filter then only approximates the optimal latent-space response.

A promising extension is to relax these assumptions by estimating per-timestep spectra, designing content-aware filters directly in the latent space, and augmenting them with lightweight nonlinear corrections, while preserving the plug-and-play nature of our cache policy. These extensions would reduce the gap between the assumed and actual signal models and further improve fidelity under real-world deviations from our assumptions.

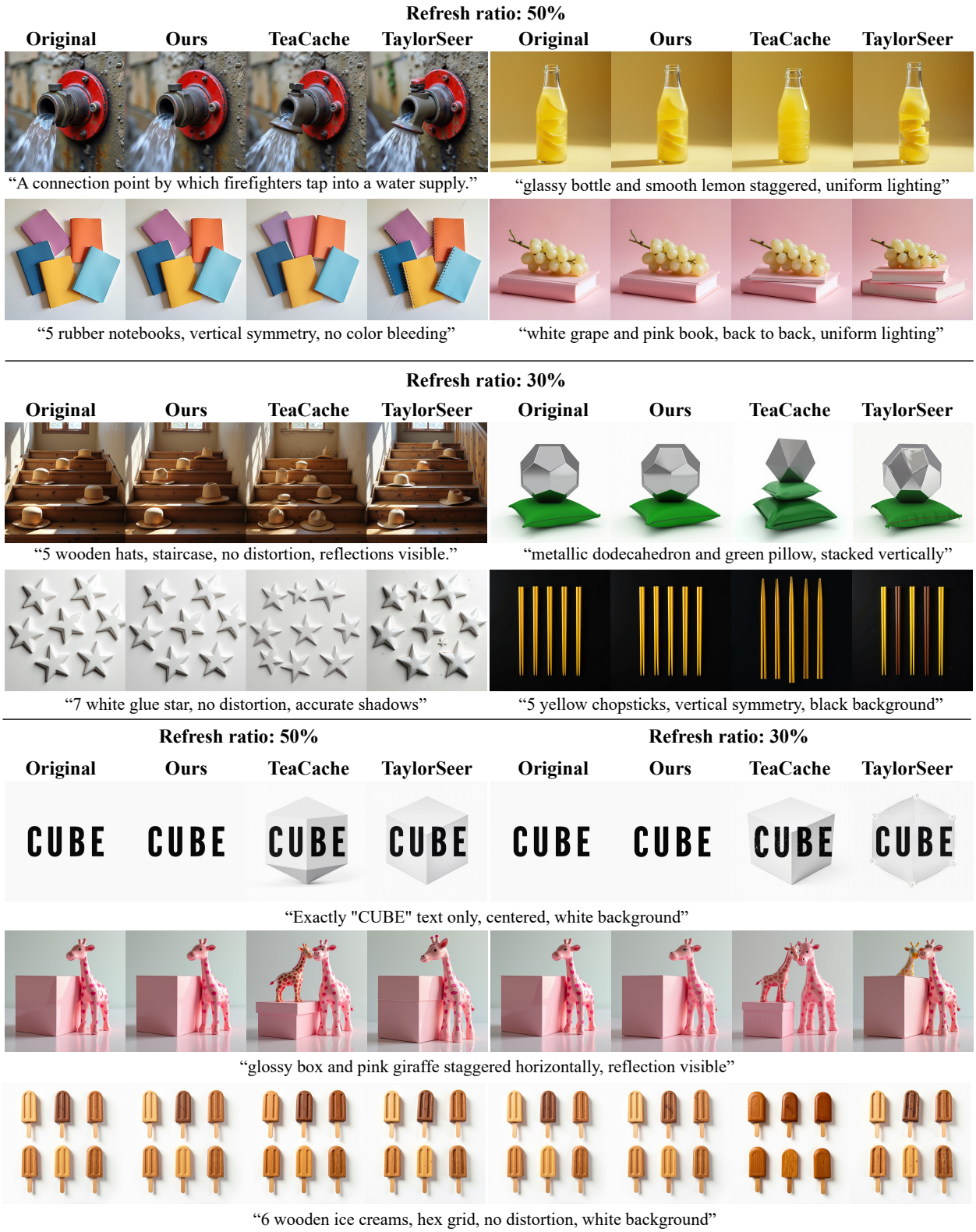

Figure 1. Additional qualitative comparison of SeaCache and baselines on *FLUX* at refresh ratios of approximately 30% and 50%.

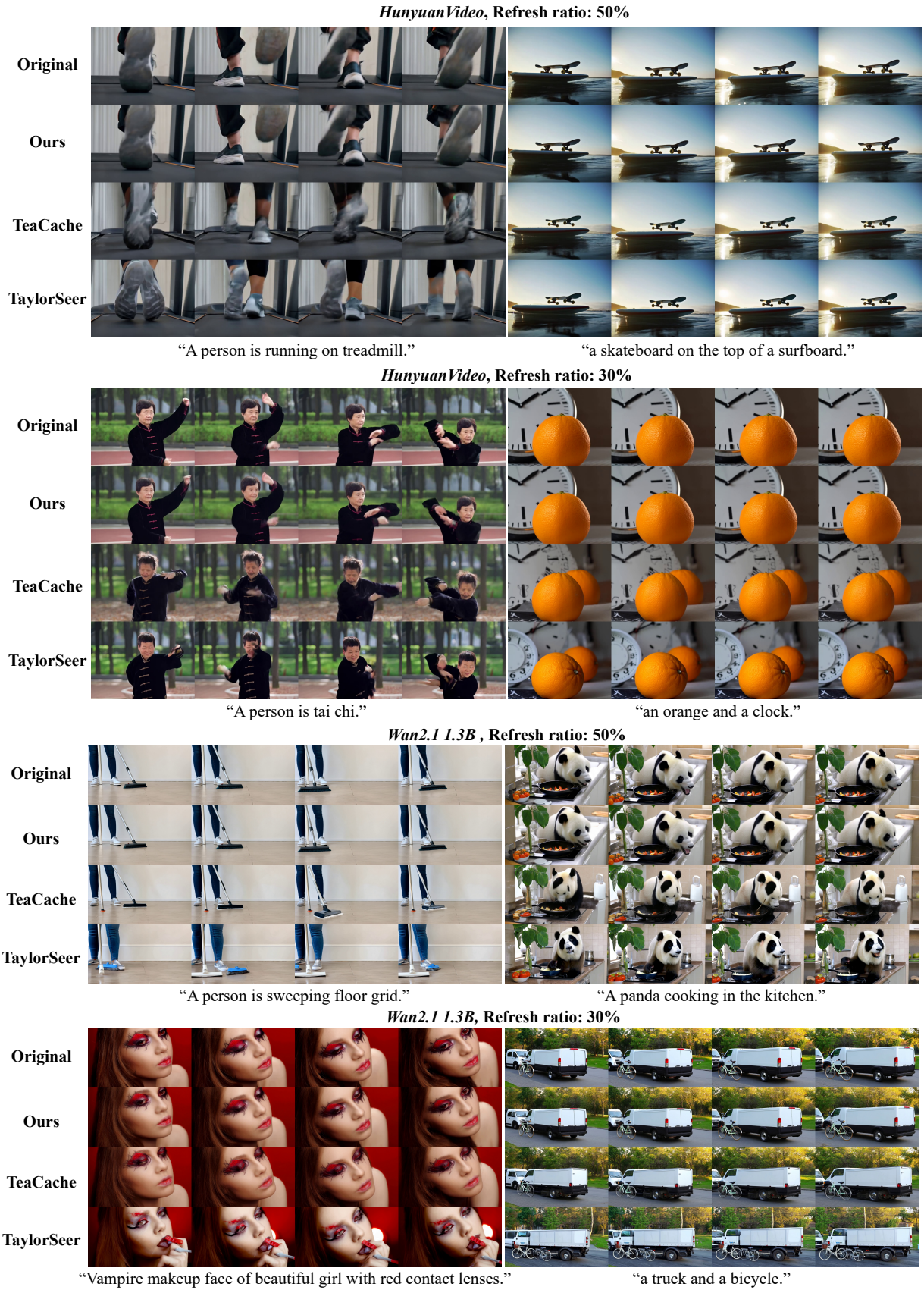

Figure 2. Additional T2V qualitative comparison of SeaCache and baselines at refresh ratios of approximately 30% and 50%.

References

[1] David H Brandwood. A complex gradient operator and its application in adaptive array theory. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, pages 11–16. IET, 1983. 1

[2] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987. 2

[3] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. 2

[4] Xiaoliu Guan, Lielin Jiang, Hanqi Chen, Xu Zhang, Jiaxing Yan, Guanzhong Wang, Yi Liu, Zetao Zhang, and Yu Wu. Forecasting when to forecast: Accelerating diffusion models with confidence-gated taylor. *Knowledge-Based Systems*, page 114635, 2025. 2

[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[6] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[8] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[9] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7353–7363, 2025. 2

[10] Stephane Mallat. *A wavelet tour of signal processing*. Academic press, 1999. 1

[11] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 2, 4

[12] David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 2

[13] van A Van der Schaaf and JH van van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996. 2

[14] 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*, 2025. 2

[15] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*. The MIT press, 1964. 1

[16] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023. 1