

# DYNAMICEval: RETHINKING EVALUATION FOR DYNAMIC TEXT-TO-VIDEO SYNTHESIS

**Anonymous authors**

Paper under double-blind review

## A VISUAL EXAMPLES

Please note that we have provided an HTML page containing different visualizations supporting our results. Extract the zip file and run `index.html`. We will refer to some of these visualizations in the subsequent sections.

## B DATASET CURATION AND SUBJECTIVE STUDY

### B.1 PROMPT CURATION

The prompts are generated based on three key aspects: background scene, object/subject, and camera attributes. We collect various classes of these key aspects from existing databases to generate the text prompts. We describe the details of this data collection:

1. **Background Scene:** We use the Places365 Zhou et al. (2017) dataset to obtain a list of 434 background scenes and manually classify them into indoor, outdoor-land and outdoor-water. This classification helps pair the other key aspects in a realistic manner with the relevant background scenes.
2. **Object/Subject:** To have a primary object/subject of focus, we collect 80 categories of objects described in MS-COCO dataset Lin et al. (2014) in which 19 are human/animal subjects while others are inanimate objects. Further, we collect 278 human/animal/vehicle subjects from ComCLIP dataset Jiang et al. (2024).
  - (a) **Subject Action:** From Kinetics700 Carreira et al. (2019) we first follow a graph based algorithm Hongjin et al. (2022) to select 100 semantically diverse human actions and pair them with human subjects. The semantic diversity is achieved by clustering the sentence embeddings of the human action phrases and equally sampling from each cluster. ComCLIP Jiang et al. (2024) also provides human/animal/vehicle actions that are paired with relevant subject categories along with the scene classification (indoor/outdoor).
  - (b) **Number of Subjects:** Many generative models fail to keep the subject shape consistent if there are more than one subjects in the scene, causing merging/splitting of subjects. To enable evaluation of such issues we add a subject attribute that specifies the number of subjects in the scene ('one', 'two', 'three' or 'many').
3. **Camera Attributes:** The crucial part of our benchmark is to have camera motion that helps evaluate the model's ability to generate good quality dynamic videos. There is no existing work that explicitly provides a list of camera types and motions. Thus, we extract such keywords from prompt benchmark datasets Bain et al. (2021); Liu et al. (2024b); Huang et al. (2024) manually and classify them into camera type and motion often paired together and classified into outdoor/indoor.
  - (a) **Camera Type:** The camera type describes the initial camera setting while capturing a video. It could describe the focal length, positioning or type of the camera (wide angle, medium shot, aerial shot, low angle shot or helicopter camera). We collect 15 camera types from the prompt benchmarks.
  - (b) **Camera Motion:** To introduce camera motion, we collect a list of 26 diverse camera motions (smooth dolly move, arc shot, trucking shot or follow subject).

We randomly sample each element and its paired attributes to construct a scene metadata in JSON format and ask GPT-4o to generate a plausible description of the scene. An example of generated prompt from such metadata is shown in Tab. 1.

Table 1: Structured metadata (left) and its descriptive narrative (right).

Scene meta-data	Generated prompt
<pre>{   "setting": "indoor",   "action_dataset": "comclip",   "metadata": {     "scene": "auto factory",     "subject": {       "name": "dog",       "number_of_subjects": "one",       "action": "drinking the water"     },     "camera": {       "type": "ground shot",       "movement": "dolly shot"     },     "extra attributes": ""   }, }</pre>	<p>"In an auto factory, a lone dog drinks water from a puddle amidst hulking machinery and assembly lines. The ground shot dolly camera moves forward, revealing industrial surroundings with the dog in sharp focus."</p>

## B.2 EXAMPLES FOR EACH EVALUATION DIMENSIONS

We have provided examples of video pairs and their ground truth scores in "Pairwise Comparisons" section in our HTML page. For convenience we have arranged the video pairs such that the video on the left has higher quality. Most commonly background scene consistency gets affected by localized distortions in the background scene, and for foreground object consistency, humans prefer rigid objects over morphing.

## B.3 VIDEO GENERATION AND CONSTRUCTING PAIRS

We have provided a grid of generated videos for some of the prompts in "Dataset Videos" section in the HTML page. The video grid is arranged such that each column represents a different model and each row in the column contains different random generations from the model. Most of the videos contain significant camera motion.

We conduct a subjective study where we provide a pair of videos generated from the same prompt and ask the human subjects to select the video based on each evaluation dimension. Given a prompt that has  $10 \times 3 = 30$  videos, we generate a controlled number of pairs that compare videos generated from the same model and across models, instead of evaluating on all  ${}^{30}C_2$  pairs. All combination of pairs from the same model are compared generating  ${}^3C_2 = 3$  pairs per model per prompt,  ${}^3C_2 \times 10 \times 100 = 3,000$  intra-model pairs. For inter-model comparisons, for each prompt, one video is randomly selected from each model. All the combinations of pairs across the selected videos constitute  ${}^{10}C_2 = 45$  inter-model pairs per prompt,  ${}^{10}C_2 \times 100 = 4,500$  inter-model pairs in total. Combining both, we evaluate  $3,000 + 4,500 = 7,500$  video pairs in total.

Each video pair is evaluated across the 2 dimensions. Thus, there are  $((2 \times 100) \times (30 + 45)) = 15,000$  evaluation pairs in total. We collect 3 subjective ratings per pair totaling  $3 \times 15,000 = 45,000$  human ratings. We employ various levels of reliability checks to ensure the quality of annotated data.

## B.4 SUBJECTIVE STUDY

**Qualification Test:** First, a pool of workers are presented with a qualification test in which they are provided an instruction video that outlines the details of our study with examples shown for each dimension. The qualification test checks if the workers understand the instructions by asking 10 multiple choice questions (MCQs). Further, we provide 3 gold standard video pairs and ask the workers to select the correct video from each pair for specific evaluation dimensions. The workers are qualified for the main study if they obtain a qualification test score greater than 8 out of 10 and select the correct video in all the 3 pairs. Additionally, we filter the qualified workers on their study setup (screen resolution, size, device). Through the qualification test we select 65/200 workers for the main study.

**Main Subjective Study:** The main subjective study is designed with more internal reliability checks. In each Human Intelligence Task (HIT), we collect ratings on 15 video pairs, and each video pair is evaluated on all dimensions, totaling 45 pair evaluations on average. We employ 3 new reliability checks involving repeated questions, gold standard pairs and video level sanity checks. For repeated questions, we select 2 out of 15 pairs and repeat them in the HIT after shuffling the video pair and the evaluation dimensions. We manually collect a gold standard set that contains annotations for pairs that are easy to differentiate. We add two random 2 gold standard pairs that contains 2-3 questions. Finally, we manually collect a set of sanity check questions (MCQs) about the content in a video pair; per HIT, we include 1 sanity check that contains two questions. Thus, there are  $15 + (2 + 2 + 1) = 15 + 5^{\text{ReliabilityChecks}} = 20$  video pairs in total in one HIT. In the 5 video pairs used for reliability checks, there are around 12 to 14 questions. The HITs are approved if the worker clears atleast 80% of the reliability questions. With all these checks in place, we collect a highly reliable set of annotations for all the video pairs.

**Worker Compensation:** The worker compensation is fixed based on the US federal laws for a minimum wage of \$7.5 per hour. The qualification test takes around 8 minutes. Setting \$7.5 per hour, each approved qualification test is compensated with \$1. In the main study, one HIT takes around 35 minutes. Setting a wage of \$8 per hour, each approved worker is compensated with \$5.

### B.4.1 USER INTERFACE:

We design a simple user interface for an HIT with multiple pages. Each page contains a video pair and two questions (evaluation dimensions) as shown in Fig. 1. Both videos are set to autoplay on repeat for convenience. The users can make them full screen if required. We provide short instructions on the left and a separate page with detailed instructions, which contains an instruction video used in the qualification study, with examples detailing how to select a video for each dimension. The human subject must select the video with more distortion with respect to the evaluation dimension. Even if both videos look equally distorted, the subjects are prompted to re-watch and find subtle differences. The workers are allowed to submit the annotations only if all questions from all pairs are answered in the HIT.

## B.5 FOREGROUND OBJECT CONSISTENCY ON VIDEO PAIRS WITHOUT OBJECTS

Despite providing foreground object descriptions, some T2V models (mostly older or smaller models) still fail to generate scenes that include the specified objects. We conducted a manual check and observed that out of all the generated videos, around 4% of the videos fail to show the primary objects. Further, after constructing video pairs, only 1% of the pairs contained both videos without object generation, while in 6% of the pairs, one of the videos failed to generate objects. We discard the 1% cases where both videos fail to generate foreground objects. In the 6% of the cases where one video failed to generate the object, that particular video is automatically preferred less, as the model was incapable of generating the foreground object. For the subjective study, we only provide the remaining 94% of the pairs for human evaluation.

## B.6 ADDITIONAL DATA COLLECTION

Adding to the two key aspects of background scene consistency and foreground object/subject consistency, we have collected data on more evaluation dimensions for future work:

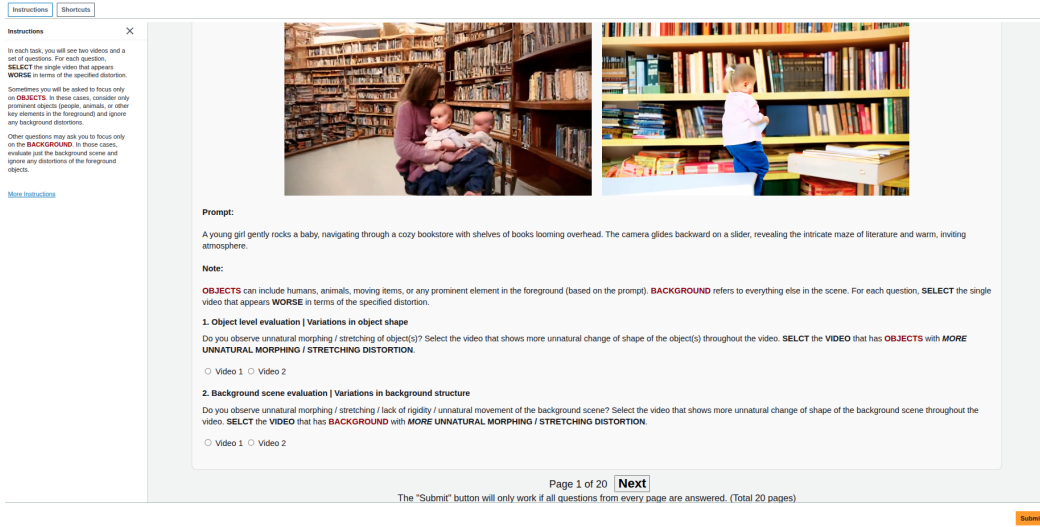


Figure 1: A screenshot of the User Interface for pairwise comparison

1. **Foreground object/subject consistency:** A poorly generated video can have foreground objects unnatural changes in their shape and size across the video, irrespective of the background scene being consistent. The object level evaluation can be divided into three dimensions, namely:
  - (a) **Object shape variation:** This dimension evaluates unnatural morphing, stretching, merging or splitting of foreground objects/subjects in a scene. For example, under camera motion, a generative model may fail to keep the shape of a table in a scene rigid.
  - (b) **Relative object size variation:** In some cases, even if the shape of the object is consistent across frames, the relative size of the object with respect to the background scene may change unnaturally when there is a camera motion. For example, the generative model fails to keep the rate of change of object size proportional to the speed of camera movement.
  - (c) **Unnatural object size:** The generative model can also fail to capture the real life proportions of different objects in a scene, irrespective of objects being consistent in shape or size across the video. One may be able to assess the size unnaturalness at a frame level, however, unless the camera moves, one cannot position an object in a scene to be closer or farther from the camera. Thus, this dimension must be evaluated at a video level.
2. **Background scene evaluation:** Regardless of whether the objects in the video have consistent shape, the background scene can have unnatural variations affecting the perceptual quality. The background consistency aspect is evaluated on two factors:
  - (a) **Background scene shape variations:** Similar to object shape variations, the background scene in a generated video also can undergo unnatural morphing or stretching with camera movement.
  - (b) **Changing scene with moving camera:** In cases where the camera moves away and returns to the same spot, the entire scene can change, as some generative models only focus on generating the next frames without keeping memory of the scene that was already generated. Even if the background is rigid, unexpected scene change can affect the viewing experience. This dimension is only evaluated if the prompt mentions such a camera motion.
3. **Object-background interaction:** Independent evaluation of foreground objects and background scene can overlook the factors that ground the objects to the actual scene. Thus, we need to evaluate how well the objects interact with the background scene.

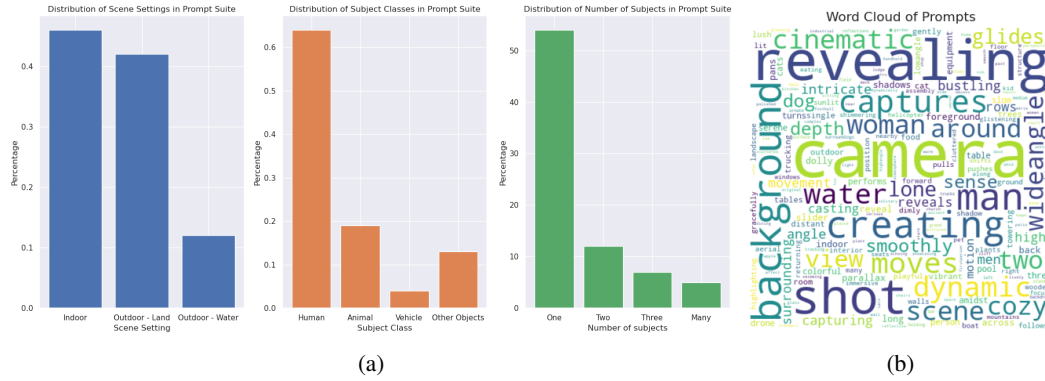


Figure 2: Dataset analysis: Percentage of categories of scene elements in the prompts suite, and the prompt wordcloud. A majority of the words focus towards camera movement.

- (a) **Sliding:** Generative models often find it difficult to keep an object fixed to the ground or platform on which it should be placed, especially when it needs to capture camera motion. Objects tend to slide off the ground, causing humans to prefer them less.
  - (b) **Shadow inconsistencies:** The shadows cast by objects must conform to the laws of physics, particularly, should be cast based on the position of the light source and should remain consistent with the shape and motion of the object no matter how the camera moves.
  - (c) **Reflection inconsistencies:** Similar to shadow evaluation, reflective surfaces can also be evaluated on how natural the reflections look and how well they reflect nearby objects. Generating reflections under camera motion is more challenging for a generative model, as it needs to regenerate a different view point of a nearby object in good detail.
- Note:** The dimensions of shadow and reflection consistency are only evaluated when the prompt contains a mention of shadow or reflection.

#### 4. Camera motion evaluation: With all the aspects visible in the scene addressed, the last quality aspect that remains to be evaluated is the camera motion itself.

- (a) **Unpleasant camera movement:** Videos with unpleasant camera motion can affect the perceptual quality of the video. We ask the subjects to evaluate how unpleasant the perceived camera motion feels.

Finally we evaluate on an additional dimension that collects a holistic human preference, given the prompt and a pair of videos. In a holistic evaluation, one would select the most preferred video considering various aspects like visual quality, motion quality, overall appeal and text-alignment. Holistic preference data helps in evaluating which individual dimensions contribute more to the final preference.

## C DATASET ANALYSIS

### C.1 PROMPT DISTRIBUTION

We show the percentage of broad categories in each scene element used to construct the prompt suite in Fig. 2a. In scene setting, there three categories: indoor, outdoor-land, outdoor-water; subjects are broadly classified into human, animal, vehicle, and other objects; for number of subjects, we sample from one, two, three, and many. Additionally, we also show the wordcloud generated from our prompt suite in Fig. 2b. We observe that a majority of the words focus primarily on camera motion and dynamic scenes.

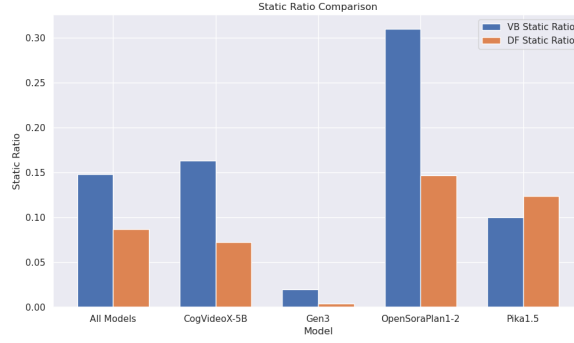


Figure 3: Comparison of ratio static videos in DynamicEval (DE) vs VBench (VB) dataset on matching models. The ratio of static scenes generated from VBench prompts are significantly higher than the ones generated from DynamicEval.

## C.2 RATIO OF DYNAMIC SCENES IN DYNAMIC EVAL VS VBENCH

### C.2.1 CAMERA-MOTION METRIC

To identify static and dynamic scenes, we define camera-motion metric using point tracks from CoTracker Karaev et al. (2024). We compute point trackers on the entire video using CoTracker and find the variance of each point tracker across frames. All the variances are averaged across points as the final camera-motion metric. Let all the point trackers be denoted as  $\{T_p^f\}_{p=1}^P$  where  $f \in \{1, 2, \dots, F\}$  represents the frame number. The camera-motion metric is computed as:

$$C_{\text{cam}} = \frac{1}{P} \sum_{p=1}^P \text{Var} \{T_p^f\}_{f=1}^F \quad (1)$$

Higher  $C_{\text{cam}}$ , higher the camera motion. We do not explicitly mask out foreground objects for this metric. Even a slight camera motion causes all trackers to shift, resulting in a much higher camera motion metric. When the camera is static, most tracks remain stationary even if foreground objects move. Thus, we compute the camera motion metric on all trackers. We find  $\tau_{\text{cam}}$ , the threshold that differentiates static vs dynamic by finding the threshold at which the 10% of videos with lowest  $C_{\text{cam}}$  are lesser than  $\tau_{\text{cam}}$ . This selection based on the observation mentioned in Section 3.3 of the main paper.

### C.2.2 COMPARING WITH VBENCH

We use the camera-motion metric to compare the ratio of static scenes generated from the prompts in VBench with the scenes generated from our DynamicEval. We utilize the videos made available by the VBench authors for CogVideoX, OpenSoraPlan, Gen3 and Pika. We randomly sample unique prompts from the VBench prompt suite and extract the corresponding videos for each model. We compute  $C_{\text{cam}}$  on these videos and videos in DynamicEval corresponding to the selected models. We partition each subset into static and dynamic using  $\tau_{\text{cam}}$ . The ratio of static scenes in both databases across models is shown in Fig. 3. The ratio of static scenes in VBench is significantly higher than the ratio of static scenes in DynamicEval, validating the effectiveness of our prompt suite in generating dynamic videos.

## D ANALYSIS OF BASELINE METRICS

### D.1 BIAS OF FEATURE LEVEL METRICS WITH CAMERA MOTION

Feature level consistency metrics like VBench background consistency (VB-BG) and subject consistency (VB-SC) are prone to bias from camera motion. To validate this, we analyze frame level plot of the consistency metrics and camera-motion metric in Fig 4. The frame level camera-motion metric is calculated by splitting frames into batches of 5 and computing the camera-motion metric

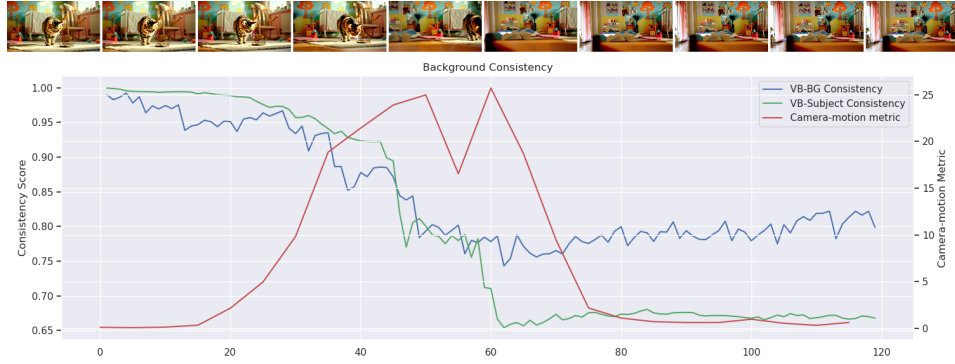


Figure 4: Effect of camera motion on feature-level metrics: We observe that with large camera motions in the scene, the feature level metrics get affected despite having high quality video.

for each batch. We select a video with large changes in camera motion and monitor VB-BG, VB-SC, and  $C_{\text{cam}}$  at a frame level. We observe that with large camera motion, the feature level metrics tend to be sensitive to such motion in the scene.

## D.2 FAILURE CASES IN BASELINE METRICS

We provide example video pairs in the supplementary HTML in "Pairwise Comparisons" section. This section shows examples where the baseline metrics fail to capture fine-grained distortions in both background and foreground. In background consistency (Click on "Background Scene Consistency" under "Pairwise Comparisons"), most of the fail cases of VB-BG correspond to videos with significantly lesser camera motion. Localized distortions are never captured as shown in Example 1 in the page. Even with severe distortions in Video 2 of Example 1, VB-BG still prefers Video 2. All the other examples also reinforce the same point. In foreground object consistency (Click on "Foreground Object Consistency" under "Pairwise Comparisons"), VB-SC also suffers from high camera motion bias as seen in Examples 1, 3 and 4. Additionally, there are cases where in case of multiple subjects in a scene VB-SC struggles to evaluate each object for reliable evaluation (Example 2).

## E PROPOSED METRICS

### E.1 IMPLEMENTATION DETAILS

#### E.1.1 MS-DEBIAS: DETAILS OF MULTI-SCALE PROCESSING

We evaluate MS-Debias metric at multiple scales using Gaussian pyramid down-sampling to create multi-scale videos before feeding them into the interpolation model. We then apply our debiased motion smoothness pipeline at each scale. To validate the effectiveness of each scale in our MS-Debias metric, we evaluate the performance of our method at different scales as shown in Table 2. We report pairwise video preference accuracy for both our method and the baseline motion smoothness metric at each scale. We observe improved performance at lower scales, consistent with prior findings that perceptual video quality is often better characterized at reduced resolutions Soundararajan & Bovik (2012). Finally, incorporating weighted multi-scale processing, our method substantially outperforms the baseline. We choose the weights for multi-scale processing proportional to the size of the bands. Specifically, we assign weights of  $1/8$ ,  $1/4$ ,  $1/2$ , and 1 to the original resolution and to videos downsampled by factors of 2, 4, and 8, respectively. We normalize these weights to sum to 1 before computing the weighted score.

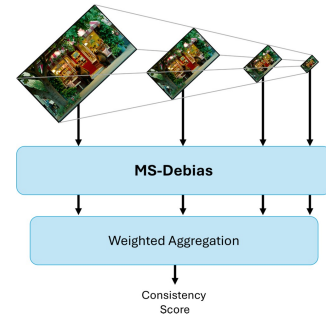


Figure 5: Multi-scale processing in MS-Debias

Table 2: Pairwise video selection accuracy of MS-Debias on DynamicEval (background scene consistency) for each scale in the Gaussian pyramid. We find that, with lower scales the performance increases gradually, and the combination of multiple scales provides the best performance.

Metrics	VB-MS	MS-Debias
Original resolution	53.7	56.6
Downscaled by 2	55.5	57.1
Downscaled by 4	55.7	57.3
Downscaled by 8	53.3	57.6
Multi-scale combination	<b>56.3</b>	<b>58.2</b>

Table 3: Time taken on average for generation and evaluation.

BG Metric	Time	FG Metric	Time
MS-Debias	15 minutes	Tracker-FG	2.2 minutes
MS-Debias-S	10 minutes	Tracker-FG-S	1 minute
Video Generation	50 minutes		

### E.1.2 TRACKER-FG: ADDITIONAL IMPLEMENTATION DETAILS

As we first detect objects using GroundingDINO (Liu et al., 2024a) and then compute Tracker-FG, there are cases where it does not detect any object in the scene. Therefore, in a given pair, if objects are not detected in both videos, we use the baseline VB-SC metric for comparison. In cases where GroundingDINO detects objects in only one video, our metric selects that particular video as high quality, as videos with no objects are considered bad quality for foreground consistency, as discussed in Section B.5.

### E.1.3 COMPUTATIONAL COST AND PRACTICALITY IN LARGE SCALE EVALUATION:

We have provided a comparison of the average time taken to generate a video using Wan2.1 and compute the evaluation metrics with a single A100 GPU in Table 3. From the table it is evident that the time taken to evaluate the videos is considerably less than the time taken to generate the video. Additionally, we replace the heavyweight models of SAM2 (Ravi et al., 2024) and CoTracker (Karaev et al., 2024) with their light-weight variants (shown with "-S" as a suffix in the table) to further reduce the time taken for evaluation. Note that in MS-Debias, there is a step that computes video segmentation maps for every objects in the scene to extract clean edges. This mainly contributes to the time cost. With newer lightweight segmentation/tracking models, one can use our framework to further improve the speed. Additionally, there are different parts in the framework that can be parallelly computed if optimized for multi-GPU settings.

## E.2 ANALYSIS

### E.2.1 METRIC VISUALIZATIONS

We provide graphical video visualizations of our method to better understand and get an intuition of our metrics in "Metric Visualization" section of our HTML page.

**MS-Debias:** We have provided visuals of all the intermediate steps of MS-Debias in the HTML pages. Initially the baseline motion smoothness error map is mostly sensitive to the occlusions/disocclusions and foreground objects. After adding object and edge maps, the debiased error map tends to show more of the localized issues in the background.

**Tracker-FG:** We provide visualization of the average of kNN tracker distance deviation across point tracks per object. We have colored the corresponding object data for convenience. Higher the inconsistency of an object, higher the average deviation. We have also provided a similar visualization for real videos. Here we see that the average deviation stays very low compared to AI videos with object inconsistencies.



### E.2.2 QUALITATIVE ANALYSIS

We qualitatively analyze our metrics in "Pairwise Comparisons" section in HTML. For background consistency, VBench background consistency (VB-BG) fails to predict the higher quality videos especially when there is lesser camera motion. VB-BG tends to reject videos with more camera motion. Whereas MS-Debias is able to prefer the correct example in such cases by focusing on the localized distortions (Refer to all examples in "Pairwise Comparisons" - "Background Scene Consistency"). The same issue is seen in VB-Subject consistency (VB-SC), with VB-SC preferring videos with lower motion more. Tracker-FG focuses on the object without getting biased by camera or object motion. It captures long term dependencies more effectively and is capable of tracking each subject individually to compute consistency, which is missing in VB-SC.

We further analyze some failure cases where both baseline and proposed metrics fail to select the higher quality video. MS-Debias tend fail when the generated video has very gradual variations in background. The flow based method is able to reconstruct the intermediate frames even though the videos look unnatural. Humans tend to prefer more natural scenes. A similar trend is seen with Tracker-FG. Humans reject unnatural objects even if they look rigid and flow smoothly. Additionally,

### E.2.3 IMPACT ON DIFFERENT TYPES OF CAMERA MOTION:

We broadly classify camera motion described in each prompt into five subsets: linear translational motion (dolly shot, tracking shot, slider move), curved translational motion (arc shot, panning around subject), handheld motion and rotational motion (camera turns left/right, pedestal shot). We then compute the performance of each subset individually as shown in the Table 4.

Table 4: Pairwise preference accuracy on different types of camera motion.

Method	Full	Linear	Curved	Handheld	Rotation
Background Scene Consistency					
VB-BG	56.0	55.2	54.1	57.9	57.1
MS-Debias	57.0	56.8	54.1	60.0	56.6
Foreground Object Consistency					
VB-SC	57.4	57.6	58.7	56.2	56.0
Tracker-FG	58.0	57.7	59.3	58.3	57.2

### E.2.4 PAIRWISE PREFERENCE ON INTER-MODEL VS INTRA-MODEL COMPARISONS

We evaluate how well our metrics distinguish between videos generated by different models and those generated by the same model by comparing the pairwise preference accuracy of the metrics on the inter- and intra-model subsets in Table 5. In both inter-model and intra-model comparisons, our proposed metrics outperform the baselines. As intra-model comparisons involve comparing videos with similar content and resolution, the accuracy is generally higher.

Table 5: Performance of the methods on inter-model vs intra-model video comparisons. The proposed metrics perform better on both comparisons, with higher performance on intra-model pairs, as the videos contain similar content.

Comparison Type	Background Consistency		Subject Consistency	
	VB-BG	MS-Debias	VB-SC	Tracker-FG
Inter-model Comparisons	54.8	<b>58.3</b>	54.3	<b>56.5</b>
Intra-model Comparisons	57.7	<b>58.2</b>	59.1	<b>60.7</b>

### E.2.5 COMPARISON WITH MODELS FINE-TUNED ON SUBJECTIVE QUALITY

VBench and EvalCrafter provides more metrics that use models fine-tuned using quality labels on real videos, such as VB-quality (Huang et al., 2024) and EC-Dover (Liu et al., 2024b). VB-quality uses the MUSIQ (Ke et al., 2021) image quality predictor which is trained on camera captured images with subjective quality annotations. EC-Dover uses the DOVER (Wu et al., 2023) video quality prediction model, which is trained on real-world distorted videos, where human subjects annotate them on perceptual quality. In contrast, our metrics are purely zero-shot with respect to quality evaluation. We compare our background consistency metric with quality pre-trained metrics proposed in VBench and EvalCrafter in Table 6. The fine-tuned metrics generally outperform the baseline metric VB-BG, with EC-Dover achieving the best results, likely due to being trained on videos. Notably, despite being a zero shot metric, MS-Debias achieves the same performance best metric fine-tuned on subjective quality labels.

Table 6: Pairwise preference accuracy compared with models trained on subjective quality.

Method	VB-Quality	EC-Dover	VB-BG	MS-Debias
Accuracy	56.5	<b>58.2</b>	56.0	<b>58.2</b>

### E.2.6 MUTUAL OVERLAP BETWEEN BOTH METRICS

In the background scene consistency metric MSDebias, we provide an object mask to debias the effect of foreground objects in the computation. Similarly, for the foreground consistency metric Tracker-FG, we explicitly compute the tracks only in the object mask regions. Therefore, in both the metrics there is clearly no overlap in terms of the features they evaluate. In contrast, the baseline VBench metrics do not explicitly separate out the foreground and background regions as they are computed on holistic deep features. Nevertheless, to validate that there is no overlap in performance, we evaluate the pairwise preference accuracy of MS-Debias on foreground consistency and Tracker-FG on background consistency, swapping the metrics, as shown in the Table7. The poor performance of each measure in the last row indicates how well separated the metrics are.

## F MISCELLANEOUS

### F.1 EVALUATION METRICS

**Top-k video selection accuracy:** Top-k video selection accuracy is the proportion of times the highest quality video is among the top-k predicted videos from the metric. For Top-k evaluation, we obtain ground truth ranking of videos given a prompt through win ratios. To obtain the ground truth ranking of videos in a prompt, we first filter out the pairwise comparisons between all combinations of 10 videos each selected from each model to obtain  $^{10}C_2 = 45$  pairs. For each video, there will be  $10 - 1 = 9$  pair comparisons from which we compute the win ratio of the video as the number of times the video gets selected among 9 pairs. The video with the highest win ratio is considered as the highest preferred video.

Method	Full Dataset	Full Agreement
Background Consistency (Pairwise Acc.)		
VB-BG	56.0	59.3
MS-Debias	<b>58.2</b>	<b>62.7</b>
Tracker-FG*	52.7	54.4
Subject Consistency (Pairwise Acc.)		
VB-SC	56.2	58.8
Tracker-FG	<b>58.2</b>	<b>62.7</b>
MS-Debias*	53.7	53.7

Table 7: Pairwise preference accuracy after swapping the metrics

Note that we do not report Spearman’s rank order correlation, Pearson’s linear correlation or Kendall’s rank correlation at a video level, as we have collected a prompt conditioned pairwise video comparison database and not a database with absolute score per video. Rank correlations at a video level can only be applied to an overall video level score that is not conditioned on any variable (eg. prompt).

## G BROADER IMPACT AND LIMITATIONS

DynamicEval enables research on fine-grained evaluation of text-to-video models ensuring human preference alignment on a video level. This promotes development of metrics that correlates well with human preference not only to select better T2V models, but also to select better videos given a pool of candidate videos from the best T2V models, leading to high-quality video content in real-world applications. Our interpretable design of metrics enables researchers to localize and mitigate quality issues in generated videos. Better evaluation tools can lead to higher quality content generation, which can be misused for malicious purposes (generating realistic fake videos). This raises the need for responsible deployment and regulations.

In addition to risks from generative model misuse, we identify two potential scenarios where our evaluation metrics could themselves be misused. First, if our metrics are used for real/fake detection, they may incorrectly classify high-quality generated videos as real. We therefore recommend using these metrics in conjunction with established real/fake detectors. Another potential for misuse is in evaluating and selecting generative models based on these metrics, that can be prone to over-optimization. Therefore, it is always ideal to mix multiple metrics (and potentially keep adding newer metrics) for a more holistic evaluation of the generated videos.

## REFERENCES

- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.
- Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.
- SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2022.
- 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*, pp. 21807–21818, 2024.
- Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. Comclip: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6639–6659, 2024.
- Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2024.
- Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024a.
- 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024b.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- Rajiv Soundararajan and Alan C Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2012.
- Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023.
- Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.