Rethinking Human Evaluation Protocol for Text-to-Video Models: Enhancing Reliability, Reproducibility, and Practicality

Tianle Zhang, Langtian Ma, Yuchen Yan, Yuchen Zhang, Kai Wang, Yue Yang, Ziyao Guo, Wenqi Shao, Yang You, Yu Qiao, Ping Luo, Kaipeng Zhang

NeurIPS 2024

Text-to-video (T2V) technology & human evaluation

Table 6: Full list of surveyed papers, where - indicates not mentioned in the article.

Title	Humeval	Validity	Crowds	Annotators	Training	Format	Venue	Year
GVSD [87]								
TGAN [75]								
Sync-DRAW [60]								
ASVGC [57]								
TGANs-C [69]								
MoCoGAN [83]								
CVGST [30]								
V2VSynthesis [90]								
FRGAN [121]								
MD-GAN [104]								
PSGAN+SCGAN [106]								
Gist [51]								
FBF+TS+FG [9]								
Few-shotV2V [89]								
Seg2Vid [68]								
IRC-GAN [16]								
TFGAN [1]								
G3AN [95]								
DTVNet [116]								
CAR-Nets [91]								
UOD [4]								
SIVS [17]								
PVG [59]								
SDTFG [19]								
GODIVA [97]								
MMVID [29]								
Imagen Video [35]								
VDM [36]								
Make-A-Video [78]								
StyleGAN-V [79]								
DIGAN [113]								
FDMLV [31]								
DCK [108]								
MMVID [29]								
Phenaki [85]								
NÜWA [99]								
NUWA-Infinity [98]								
FETV [54]								
MAGE [38]								
VideoLDM [5]								
PYoCo [24]								
LVDM [32]								
Videogen [50]								
ModelScope [88]								
Tune-A-Video [101]								

Table 7: Full list of surveyed papers, where - indicates not mentioned in the article.

Title	Humeval	Validity	Crowds	Annotators	Training	Format	Venue	Year
LAVIE [96]	✓	×	×	30	×	comparative	arXiv	2023
NUWA-XL [110]	×	×	×	-	×	-	arXiv	2023
Show-1 [114]	✓	×	✓	-	×	comparative	arXiv	2023
MotionDirector [122]	✓	×	✓	5	✓	comparative	arXiv	2023
MagicVideo [123]	✓	×	×	-	×	comparative	arXiv	2023
VideoCrafter1 [10]	✓	×	×	-	×	absolute	arXiv	2023
SadTalker [118]	✓	×	×	20	×	comparative	CVPR	2023
Gen1 [20]	✓	×	✓	5	×	comparative	ICCV	2023
Text2Performer [42]	✓	×	×	20	×	absolute	ICCV	2023
Text2Video-Zero [45]	×	×	×	-	×	-	ICCV	2023
VideoFusion [55]	×	×	×	-	×	-	CVPR	2023
DynamiCrafter [102]	✓	×	×	49	✓	comparative	arXiv	2023
MCDiff [12]	×	×	×	-	×	-	arXiv	2023
DragNUWA [109]	×	×	×	-	×	-	arXiv	2023
Control-A-Video [13]	✓	×	×	18	×	absolute	arXiv	2023
DreamPose [44]	✓	×	✓	50	×	absolute	ICCV	2023
VideoComposer [93]	×	×	×	-	×	-	arXiv	2023
MagicAvatar [115]	×	×	×	-	×	-	arXiv	2023
Emu Video [25]	✓	×	×	5	✓	comparative	arXiv	2023
MAGVIT [111]	×	×	-	-	×	-	CVPR	2023
I2VGen-XL [117]	×	×	-	-	×	-	arXiv	2023
SVD [3]	✓	×	-	-	×	comparative	arXiv	2023
LFDM [63]	×	×	-	-	×	-	CVPR	2023
MAGE [38]	✓	×	×	16	×	absolute	TMM	2023
CogVideo [37]	✓	×	×	90	×	absolute	ICLR	2023
Dreamix [62]	✓	×	×	10	×	absolute	arXiv	2023
ED-T2V [52]	×	×	-	-	×	-	IJCNN	2023
Free-Bloom [39]	✓	×	×	80	×	absolute	NeurIPS	2023
MM-Diffusion [74]	✓	×	✓	-	×	absolute	CVPR	2023
PVDM [112]	×	×	-	-	×	-	CVPR	2023
VIDM [58]	×	×	-	-	×	-	AAAI	2023
EvalCrafter [53]	✓	×	×	3	✓	absolute	CVPR	2024
AIGCBench [21]	✓	×	×	42	×	comparative	TBench	2024
T2VScore [100]	✓	×	×	10	✓	absolute	arXiv	2024
VBench [40]	✓	×	×	-	✓	comparative	CVPR	2024
Seer [26]	✓	×	×	54	✓	comparative	ICLR	2024
Video Factory [92]	×	×	×	-	×	-	arXiv	2024
VideoCrafter1 [11]	✓	×	×	-	×	comparative	arXiv	2024
AnimateDiff [27]	✓	×	×	-	✓	comparative	ICLR	2024
SEINE [14]	✓	×	×	10	×	comparative	ICLR	2024
ControlVideo [119]	✓	×	×	5	×	comparative	ICLR	2024
PIA [120]	✓	×	-	-	×	comparative	CVPR	2024
SimDA [103]	✓	×	-	-	×	comparative	CVPR	2024
PEEKABOO [41]	×	×	-	-	×	-	CVPR	2024
VideoPoet [46]	✓	×	×	7	✓	comparative	ICML	2024

A large-scale review of nearly 100 articles revealed the shortcomings of existing human evaluation protocols.

Observations:

- Evaluation methods vary widely, and many lack detailed disclosure of the protocols

→ ***Lack Reproducibility***

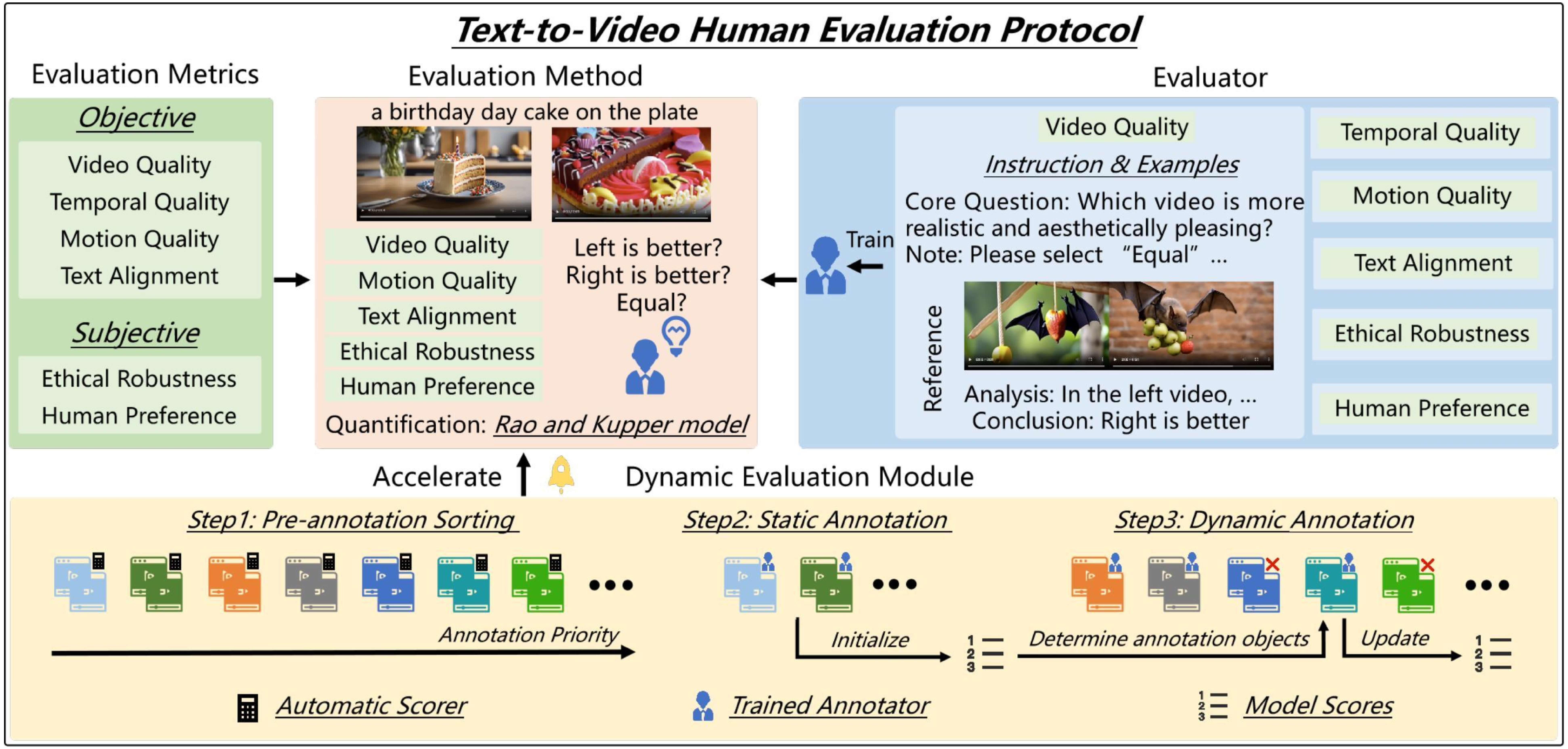
- Many employ laboratory-recruited annotators (LRAs) without corresponding training

→ ***Lack Reliability***

- Number of annotations can reach tens of thousands, and increases in $O(N^2)$ trend

→ ***Lack Practicality***

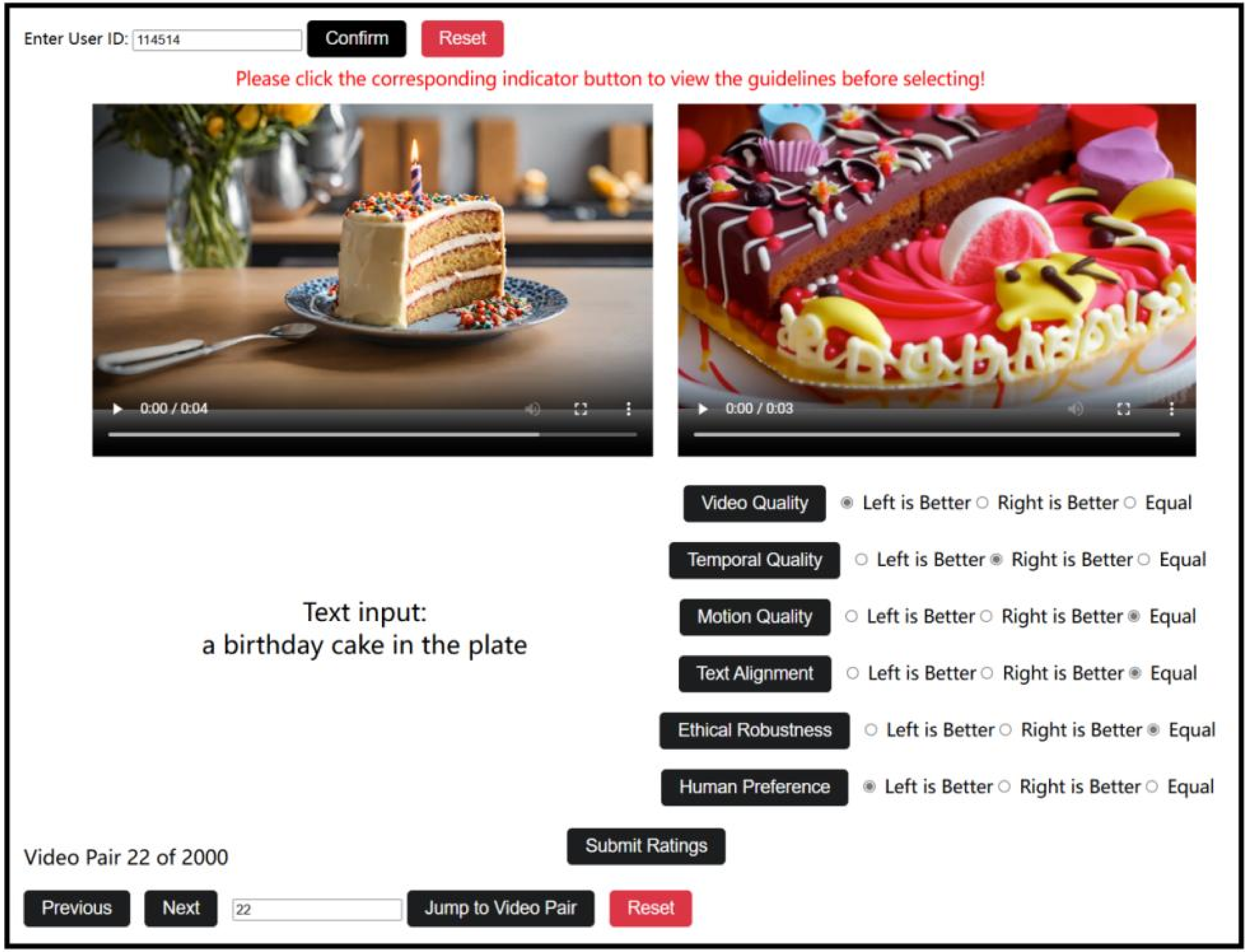
Text-to-Video Human Evaluation Protocol (T2VHE)



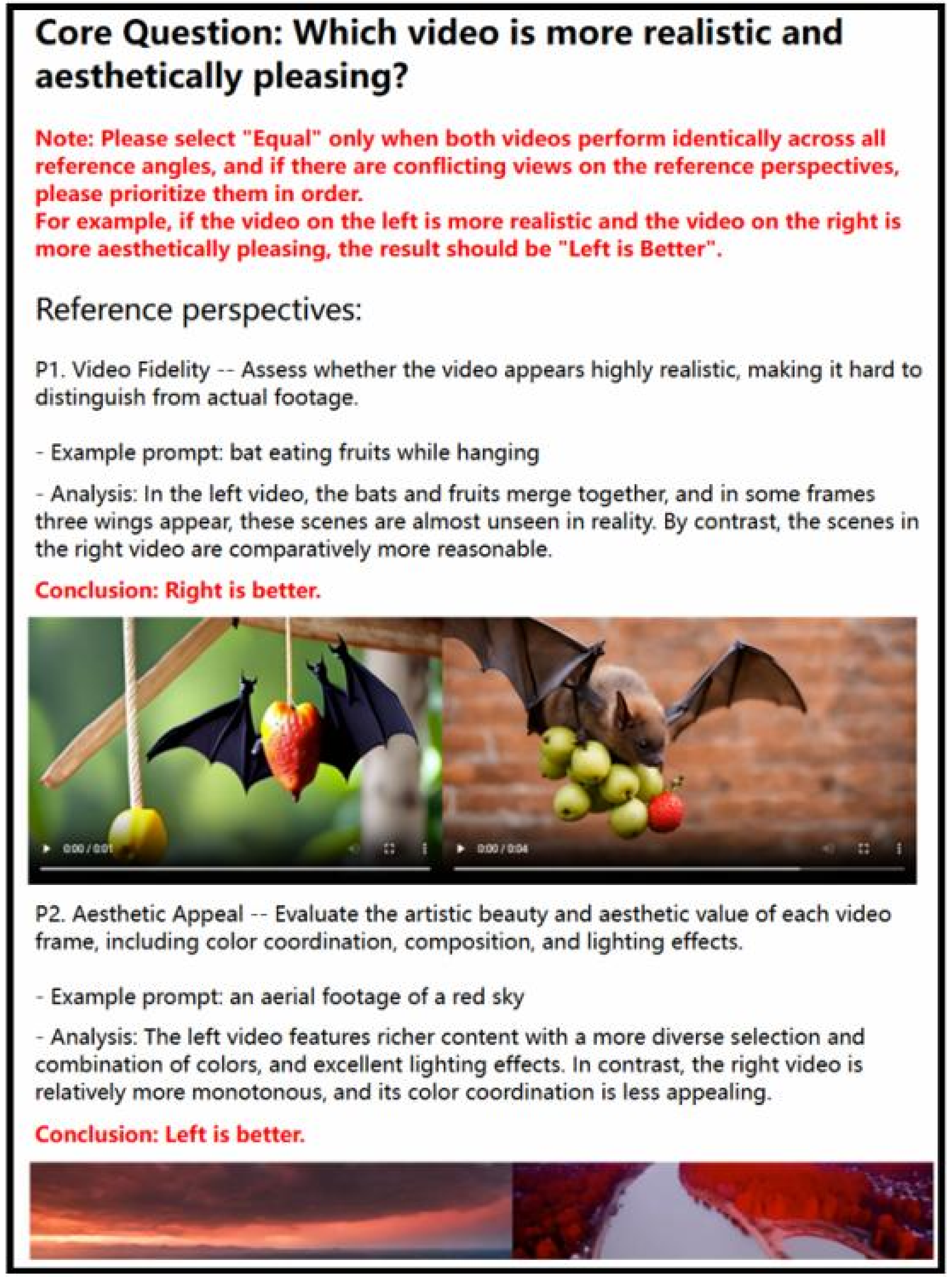
- **Evaluation metrics**: 4 objective indicators, 2 subjective indicators, each with 2 reference perspectives
- **Evaluation method**: comparative method, quantized by Rao and Kupper model
- **Evaluators**: provide detailed annotator training, support both crowdsourcing annotators (e.g. AMT) and LRAs

- **Dynamic evaluation module:** select annotation objects based on sample importance and model strength differences

Evaluation metrics & Evaluation method



Annotation interface



Instruction example

Evaluators & Dynamic evaluation module

Metric	AMT & Pre-training LRAs	AMT & Post-training LRAs	AMT
Video Quality	0.185	0.411	0.451
Temporal Quality	0.131	0.340	0.369
Motion Quality	0.088	0.338	0.249
Text Alignment	0.069	0.327	0.366
Ethical Robustness	-0.057	0.100	0.177
Human Preference	0.167	0.281	0.297

Comparison of the inter-annotator agreement (IAA)

- × Low consensus between the pre-training LRAs and AMT raters, two sets of model rankings are completely different
- ✓ Annotation quality of post-training LRAs is almost identical to that of crowdsourcing annotators, so as the rankings

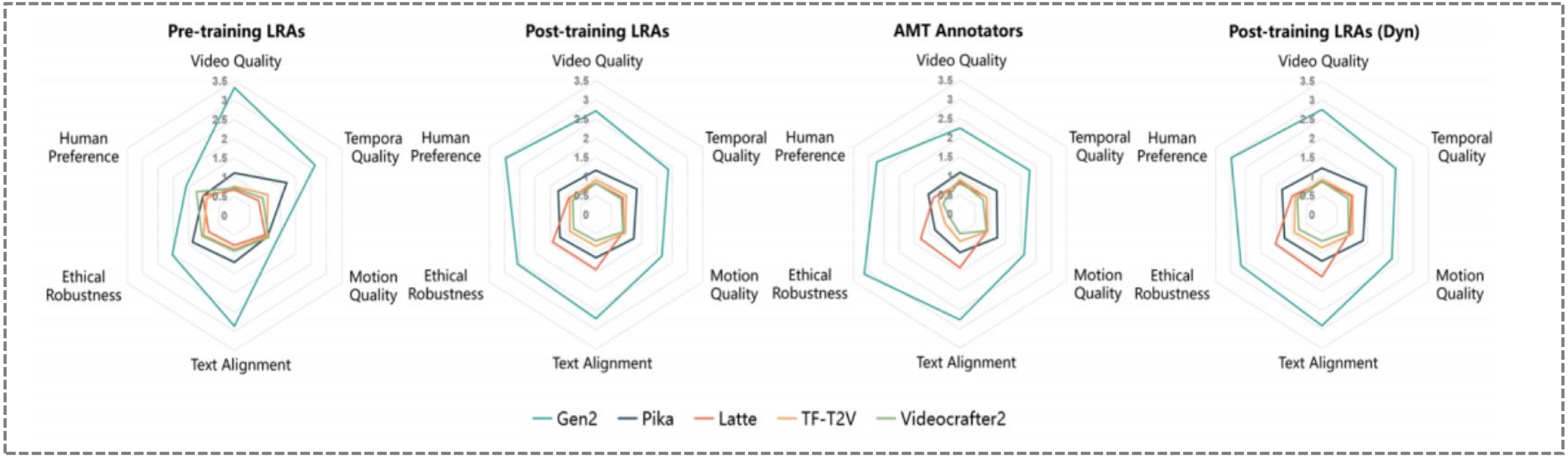
Metric	Pre-training	Post-training
Video Quality	0.224	0.339
Temporal Quality	0.178	0.288
Motion Quality	0.164	0.321
Text Alignment	0.145	0.236
Ethical Robustness	0.055	0.107
Human Preference	0.195	0.284

- ✓ Trained LRAs show significant improvement in inter-annotator agreement, i.e. annotation quality

Algorithm 1 Model Evaluation Algorithm

Input: Set of videos $\mathcal{V}$
Pre-processing:
for each video $v \in \mathcal{V}$ **do**
  compute and normalized automatic metric scores for v
  $S(v) \leftarrow$ sum of normalized scores
end for
for each prompt $\mathrm{pr}_i \in \mathcal{P}$ **do**
  for each video pair $\{v_k, v_l\} \in \mathcal{V}(\mathrm{pr}_i)$ **do**
    $pair\_score(v_k, v_l) \leftarrow f(|S(v_k) - S(v_l)|, \alpha)$
  end for
  $group\_score(\mathrm{pr}_i) \leftarrow \sum_{\{v_k, v_l\} \in \mathcal{V}(\mathrm{pr}_i)} pair\_score(v_k, v_l)$
end for
$sorted\_groups \leftarrow$ sort $\{\mathcal{V}(\mathrm{pr}_i)\}_{\mathrm{pr}_i \in \mathcal{P}}$ by $group\_score$ in descending order
Hum-evaluation:
Evaluate the first N_0 groups in $sorted\_groups$ by human and update R.
$I \leftarrow g(R)$
for each $batch$ in the remaining video pairs **do**
  for each video pair in $batch$ **do**
    Discard the video pair with probability $f(|\mathcal{F}(\{v_k, v_l\})|, \alpha)$.
    if the pair is not discarded **then**
      Evaluate the video pair by human and update R.
    end if
  end for
  $I \leftarrow g(R)$
  if model ranking is stable over 5 consecutive batches **then**
    break
  end if
end for
Output: Final model rankings and updated intensities I.

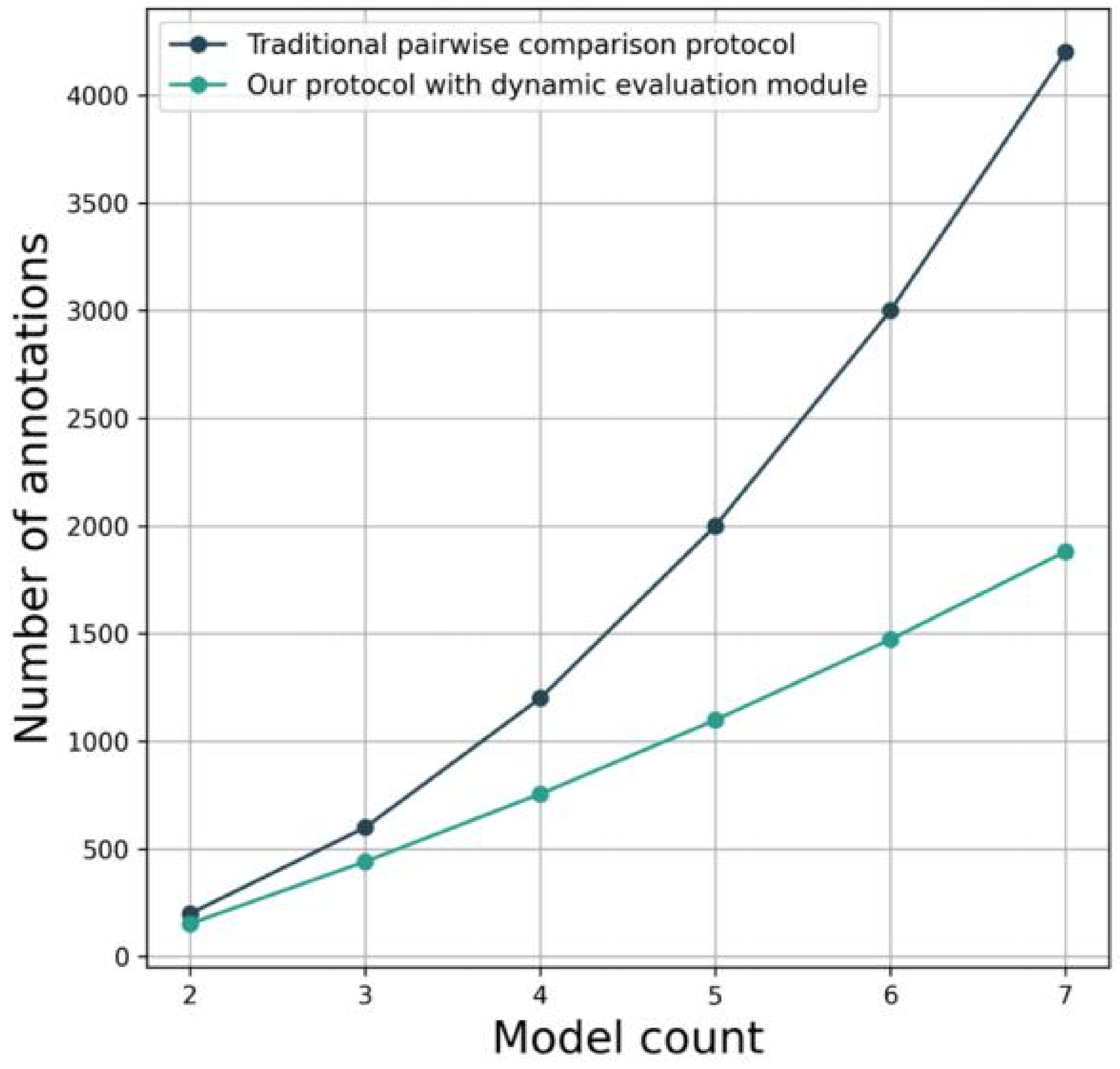
Human evaluation result

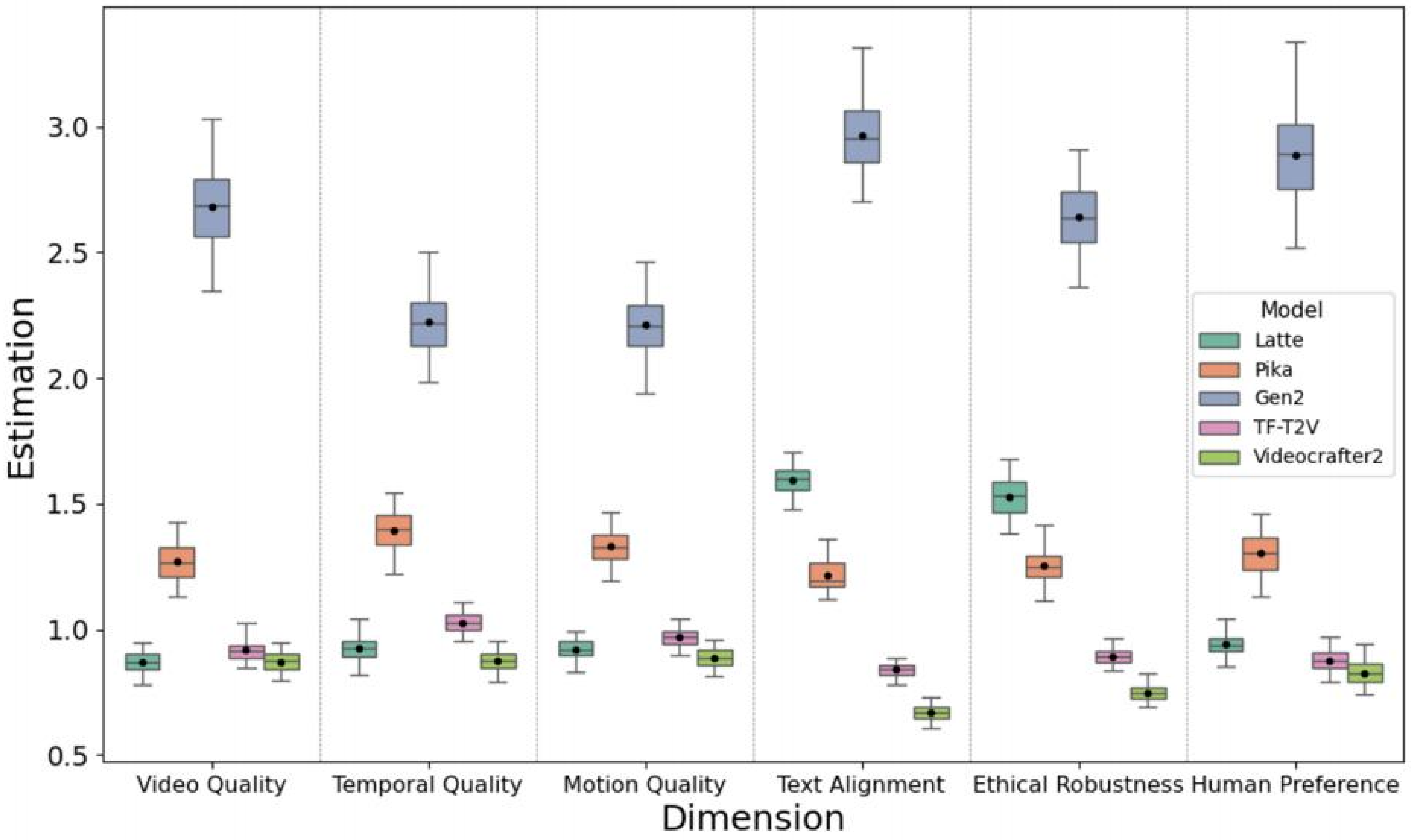


Analysis of results:

- Annotation results obtained by the pre-training LRAs markedly differ from those of the other three groups.
- Annotation results of the trained LRAs closely mirror those of the AMT personnel
- Closed-source models typically perform better.

Module validation





Verification of Effectiveness

✓ Dynamic module cuts annotation costs to about 53% of the original expense while achieving comparable outcomes.

✓ Dynamic module demonstrates a nearly linear growth in annotation demands as the number of models increases.

Verification of Reliability

✓ Pre-evaluation annotation ensures that valuable samples are not discarded

✓ Bootstrap confidence intervals shows that it only needs a small part of annotations to obtain a stable estimate of model rankings

Thank You!

Code: https://github.com/ztlmememe/T2VHE

Paper Link: https://arxiv.org/abs/2406.08845