A Evaluation Setup

We provide additional details on the evaluation setup.

Metrics. We consider the following standard video accuracy metrics: structural similarity index measure (SSIM) (Wang et al., 2004), peak signal-to-noise ratio (PSNR), learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018b), and CLIP cosine similarity score. Note that the SSIM, PSNR, and LPIPS primarily assess visual fidelity while the CLIP score captures more semantic information. We take the negative of the LPIPS score to transform it from an error metric to an accuracy metric. To compute the perceptual metrics, we resize all videos spatially to the same dimensions and subsample the longer videos to ensure that all videos have the same duration. For CLIP, we map both the ground-truth video v^{gt} and all the generated videos v_j^i to the visual-semantic space using CLIP. We compute the mean of each metric over all generated videos per task, which represents the assigned value of the metric for that task. Prior work Wan et al. (2025) on video evaluation has demonstrated the effectiveness of CLIP in capturing both visual and semantic information, motivating its inclusion in our study.

B Proofs

Proposition 1 (Uncertainty Decomposition)**.** *Define the total predictive uncertainty in the output video as the differential entropy $h(V \mid \ell)$ of the distribution $f_\theta(V \mid \ell)$. Then, this quantity can be decomposed as:*

$$h(V \mid \ell) = h(V \mid Z) + h(Z \mid \ell), \tag{3}$$

where $h(V \mid Z)$ represents the epistemic uncertainty in v, and $h(Z \mid \ell)$ the aleatoric uncertainty.

Proof. The entropy of a random variable quantifies its associated uncertainty. Given the probability distribution $f_\theta(V \mid \ell)$, we find its entropy by:

$$h(V \mid \ell) = -\int_{v \in \mathcal{V}} f_\theta(V \mid \ell) \log(f_\theta(V \mid \ell))\, \mathrm{d}v \tag{11}$$

$$= -\int_{v \in \mathcal{V}} \int_{z \in \mathcal{Z}} p(V \mid z) p(z \mid \ell) \log(p(V \mid z) p(z \mid \ell))\, \mathrm{d}z\, \mathrm{d}v, \tag{12}$$

where we incorporate the latent state generation step introduced in Equation (2). We can then decompose the $\log$ terms into two components:

$$h(V \mid \ell) = -\int_{v \in \mathcal{V}} \int_{z \in \mathcal{Z}} p(V \mid z) p(z \mid \ell) \left(\log(p(V \mid z)) + \log(p(z \mid \ell))\right)\, \mathrm{d}z\, \mathrm{d}v \tag{13}$$

$$= -\left(\int_{z \in \mathcal{Z}} p(z \mid \ell) \int_{v \in \mathcal{V}} p(V \mid z) \log(p(V \mid z)) \mathrm{d}v \mathrm{d}z \right)$$

$$- \left(\int_{z \in \mathcal{Z}} \left(\int_{v \in \mathcal{V}} p(V \mid z) \mathrm{d}v \right) p(z \mid \ell) \log(p(z \mid \ell)) \mathrm{d}z \right), \tag{14}$$

where Equation (14) applies the Fubini-Tonelli theorem. We note that each term of Equation (14) is an entropy itself:

$$h(V \mid \ell) = -\left(\int_{z \in \mathcal{Z}} p(z \mid \ell) h(V | Z = z) \mathrm{d}z \right) - \left(\int_{z \in \mathcal{Z}} p(z \mid \ell) \log(p(z \mid \ell)) \mathrm{d}z \right) \tag{15}$$

$$= h(V \mid Z) + h(Z \mid \ell). \tag{16}$$

We recognize that the first term $h(V \mid Z)$ eliminates uncertainty in prompt ambiguity, and thus signifies the epistemic uncertainty in video generation. On the other hand, the second term $h(Z \mid \ell)$ is independent of the video model, but rather only depends on the vagueness of the input prompt, signifying aleatoric uncertainty. □

C Additional Experiments

We explore the effectiveness of S-QUBED across closed-source and open-source video generation models, examining its robustness to different hyperparameters, e.g., the number of latent prompts per task and the number of generated videos per latent prompt. In particular, we consider the following state-of-the-art video models: CogVideoX Yang et al. (2024), Veo 3 DeepMind (2025), and OpenSora Peng et al. (2025). We note that these models impose different constraints that introduce practical limitations on the implementation of our method. Specifically, the closed-source model Veo 3 places a stringent cap on the number of generation tasks that a user can submit, limiting the number of latent prompts and latent videos that can be used in S-QUBED. Moreover, Veo 3 is expensive to run, introducing additional challenges. Likewise, OpenSora Peng et al. (2025) requires significant compute infrastructure with large GPU memory for video generation, e.g., an A100/H100 GPU, a potential practical bottleneck. Here, we examine if these challenges could be overcome through a more compatible choice of hyperparameters without compromising the calibration of S-QUBED. We present the additional results in the order of decreasing amount of latent prompts and latent videos. To assess calibration, we compute the correlation between the predicted uncertainty and the observed accuracy of the video model, noting that stronger negative correlation indicates better calibration of the video model. In these experiments, we use biweight midcorrelation as the correlation measure due to its robustness to outliers compared to other measures of correlation Song et al. (2012), which is especially important when analyzing data containing only a few data points. Further, we evaluate calibration on the Panda-70M dataset, given its greater diversity in the task descriptions.

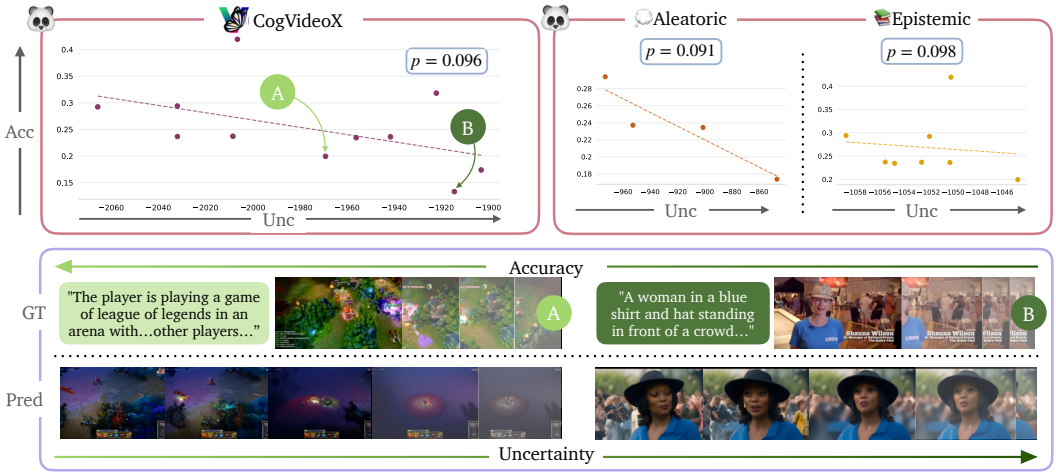

Figure 6: **Uncertainty Quantification of CogVideoX.** *Top:* S-QUBED computes calibrated uncertainty estimates for CogVideoX, decomposing the total uncertainty into trustworthy aleatoric and epistemic uncertainty components. *Bottom:* We demonstrate alignment of the uncertainty estimates with video accuracy, showing that when the video model is more uncertain, the generated video is likely to be more inaccurate, as visualized by Tasks A and B.

C.1 Results with CogVideoX

We quantify the uncertainty of CogVideoX in a random subset of 10 tasks in the Panda-70M dataset. For each task, we generate 10 latent prompts and 10 videos per latent prompt. From the resulting videos, we compute the total predictive uncertainty of the model, decomposed into its aleatoric and epistemic components. Figure 6 summarizes the quantitative and qualitative calibration results. First, as visualized in the top-left plot, we observe that the total predictive uncertainty estimates are well-calibrated. Specifically, the estimated uncertainty of the video model is negatively correlated with its accuracy, with a correlation coefficient of -0.527 at the 90% significance level. The aleatoric and epistemic estimates are also negatively correlated with accuracy at the 90% significance level, with coefficients of -0.909 and -0.625, respectively. These findings underscore that the video model's predicted uncertainty is well-aligned with the observed accuracy of the model. In other

words, when the video model is uncertain about its generated video, the synthesized video is more likely to be inaccurate. In essence, S-QUBED's uncertainty estimates are trustworthy. Further, we observe that the uncertainty estimates are well-aligned with human intuition, as demonstrated by the two video generation tasks shown in Figure 6 where S-QUBED predicts that Task A has a higher uncertainty compared to Task B. From the ground-truth and generated videos in the figure, we see that these uncertainty estimates are consistent with the relative accuracy of the video model on both tasks.

C.2 Results with Veo 3

We examine the amenability of S-QUBED to closed-source models with significant cost overhead. For example, the strict API limits on the maximum number of videos generated by Veo 3 presents notable challenges to uncertainty quantification. Likewise, each video generation task is also expensive to run. To address these challenges, we examine the calibration of S-QUBED when *fewer* latent prompts are used with *fewer* generated videos per latent prompt. Concretely, we consider five tasks, where we generate five latent prompts for each video generation task and four videos per latent prompt to estimate the uncertainty of Veo 3. We provide the calibration results in Figure 7. We find that the uncertainty estimates are strongly negatively correlated with accuracy at the 90% significance level with a correlation coefficient of -0.838. Likewise, the aleatoric uncertainty estimates are well-calibrated with a correlation coefficient of -0.937. We do not observe a statistically significant correlation between the estimated epistemic uncertainty and the accuracy of the video model, which is not surprising given that the correlation is being computed using only a few video generation tasks, making the computed correlation coefficient more susceptible to the effects of outliers. Qualitatively, as noted in our earlier discussion, the video model's uncertainty estimates are interpretable. As shown in Figure 7, Veo 3 is more uncertain about generating characters from Ni no Kuni compared to generating a video showing the Boston Celtics playing a basketball game. We see that the estimated uncertainty is well-aligned with the accuracy of the generated videos.

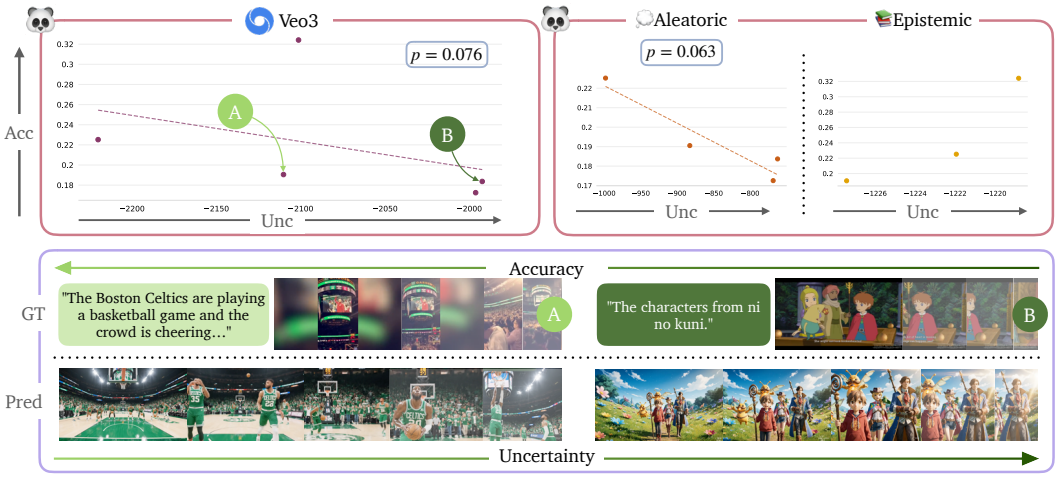

Figure 7: **Uncertainty Quantification of Veo 3.** *Top:* S-QUBED enables Veo 3 to express calibrated uncertainty estimates, although stringent limits on the number of generated videos could hinder the observation of statistically significant levels of correlation between epistemic uncertainty and accuracy. *Bottom:* The estimated uncertainty estimates are well-aligned with the accuracy of the generated videos. For example, the generated video and uncertainty estimate suggest that Veo 3 does not know the characters from Ni no Kuni compared to its knowledge of the Boston Celtics.

C.3 Results with OpenSora

We examine the applicability of S-QUBED to open-source models that require significant compute resources, e.g., models requiring large GPUs such as the A100/H100. Here, we evaluate the calibration of S-QUBED with OpenSora as a representative video model, running on an A100 GPU with 80GB memory. We estimate the uncertainty of OpenSora on 15 tasks from Panda-70M, generating three

latent prompts and three videos per latent prompt. Figure 8 shows the calibration results, highlighting that the estimated uncertainty is negatively correlated with accuracy at a 90% significance level with a coefficient of -0.481. Further, we find that S-QUBED produces well-calibrated aleatoric uncertainty estimates; however, similar to the experiments with Veo 3, we do not observe a statistically significant negative correlation between the epistemic uncertainty estimates and accuracy, likely due to the lack of sufficient generated videos to mitigate the effect of outliers. Nonetheless, the results underscore that S-QUBED produces well-calibrated total predictive uncertainty estimates, which is further demonstrated by the ground-truth and generated videos in Figure 8. The generated video in Task A is more accurate compared to the generated video in Task B, in line with the video model's uncertainty estimates.

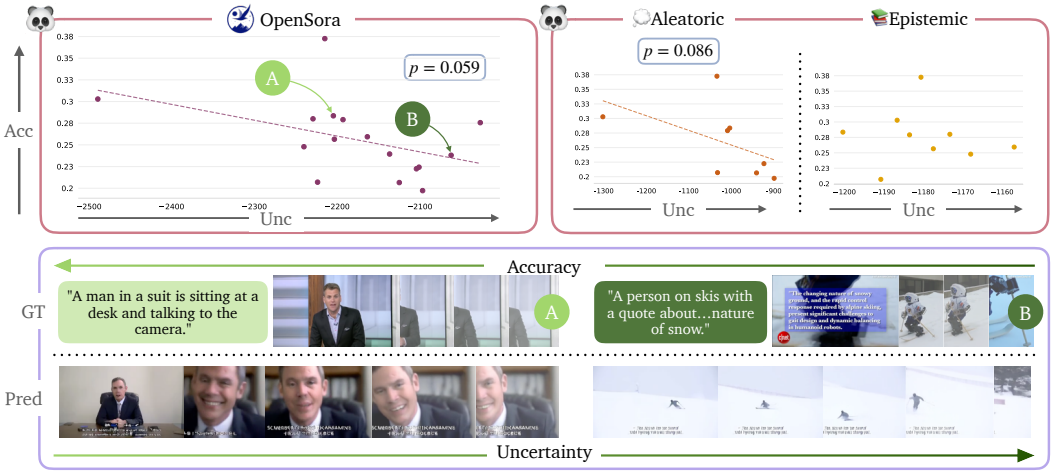

Figure 8: **Uncertainty Quantification of OpenSora.** *Top:* Empowered by S-QUBED, OpenSora produces uncertainty estimates that are well-aligned with accuracy. Unlike the aleatoric uncertainty estomates, the epistemic uncertainty does not show a statistically significant negative correlation since only a few generated videos are used in estimating the correlation. *Bottom:* The uncertainty estimates are trustworthy, with larger uncertainty estimates corresponding to greater errors between the ground-truth and generated videos.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Evan Becker and Stefano Soatto. Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441*, 2024.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008*, 2019.

Matthew Chan, Maria Molina, and Chris Metzler. Estimating epistemic and aleatoric uncertainty with a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024.

DeepMind. Veo-3: A text-to-video generation system with audio. Technical Report Tech Report, DeepMind / Google, 2025. Accessed: YYYY-MM-DD.

Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8062–8072, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Francis Galton. Note on regression and correlation. *Proceedings of the Royal Society of London*, 58: 240–242, 1895.

Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *International Conference on Machine Learning*, pp. 154–162. PMLR, 2014.

Peter Edmund Jupp and KV Mardia. A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review/Revue Internationale de Statistique*, pp. 261–294, 1989.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Igor Kononenko. Bayesian neural networks. *Biological Cybernetics*, 61(5):361–370, 1989.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Emanuele Ledda, Giorgio Fumera, and Fabio Roli. Dropout injection at test time for post hoc uncertainty quantification in neural networks. *Information Sciences*, 645:119356, 2023.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Yifei Liu, Rex Shen, and Xiaotong Shen. Novel uncertainty quantification through perturbation-assisted sample synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 46(12): 7813–7824, 2024.

Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

NVIDIA Cosmos. cosmos-predict2: General-purpose world foundation models for physical ai. https://github.com/nvidia-cosmos/cosmos-predict2, 2025. Accessed: YYYY-MM-DD.

OpenAI. GPT-5 nano, 2025. URL https://openai.com/gpt-5/. Large language model. Release date: August 7, 2025.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200k$. *arXiv preprint arXiv:2503.09642*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1):328, 2012. doi: 10.1186/1471-2105-13-328.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x). *Computational Statistics*, 27(1):177–190, 2012.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Luhuan Wu and Sinead A Williamson. Posterior uncertainty quantification in neural networks using data augmentation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3376–3384. PMLR, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.