付　　録

A.1 AutoEval-Video [17] による評価

提案手法と Video-ChatGPT [13] を AutoEval-Video により評価した結果を表 A·1，表 A·2 に示す。AutoEval-Video では，複数の能力領域やテーマにまたがるように Youtube[*1]から独自に収集され，アノテーションされたデータに対するモデルの応答の正誤をデータごとに定められた固有の評価ルールに基づいて，GPT-4 により評価した。AutoEval-Video ベンチマークでは各データは Dynamic Perception，State Transitions Perception，Comparison Reasoning，Reasoning with External Knowledge，Explanatory Reasoning，Predictive Reasoning，Description，Counterfactual Reasoning，Camera Movement Perception の九項目に分類されており，本研究では全体と各項目の正答率を評価した。AutoEval-Video は Video-based Generative Performance Benchmarking [13] とは異なり，学習データとは異なる手法で収集及びアノテーションがなされたデータに対する評価であるため，本研究ではモデルの汎化性能の評価を目的として採用した。なお，モデル構造の比較のため，別のモデルのパラメータを用いた初期化を行わず，拡張した Video Instruction Dataset [13] のみでモデル間アダプターを学習した場合の Video-ChatGPT と提案手法の結果を (W/o Init) として示している。

表 A·1　AutoEval-Video による評価結果

	All↑
Video-ChatGPT	**0.107**
Act-ChatGPT	0.058
Video-ChatGPT (w/o Init)	0.042
Act-ChatGPT (w/o Init)	0.040

表 A·2　各項目の AutoEval-Video による評価結果

	Dynamic ↑	State Transitions ↑	Comparison ↑
Video-ChatGPT	**0.089**	**0.125**	**0.211**
Act-ChatGPT	0.022	0.062	0.158
Video-ChatGPT (w/o Init)	0.042	0.094	0.158
Act-ChatGPT (w/o Init)	0.043	0.031	0.105
	External Knowledge ↑	Explanatory ↑	Predictive ↑
Video-ChatGPT	**0.088**	**0.091**	**0.125**
Act-ChatGPT	0.050	0.061	0.062
Video-ChatGPT (w/o Init)	0.012	0.030	0.031
Act-ChatGPT (w/o Init)	0.050	0.030	0.000
	Description ↑	Counterfactual ↑	Camera Movement ↑
Video-ChatGPT	**0.033**	**0.158**	0.000
Act-ChatGPT	**0.033**	0.105	0.000
Video-ChatGPT (w/o Init)	0.000	0.000	0.000
Act-ChatGPT (w/o Init)	0.000	**0.158**	0.000

表 A·1 及び表 A·2 より，AutoEval-Video による評価では提案手法はほとんどの項目で Video-ChatGPT の評価を著しく下回った。そのため，提案モデルは Video-ChatGPT よりも汎化性能が低いと言える。この結果は，それぞれのモデルに使用された学習データの量に起因すると考えられる。Video-ChatGPT のモデル間アダプターは LLaVA [4] の重みで初期化されるため，学習に使用した画像や映像の指示データの総数は 753k+100k=853k である。それに対して，提案手法のモデル間アダプターは，特に行動特徴変換モジュールと特徴変換モジュールの学習に使用された指示データの総数はデータ拡張を考慮しても 200k のみである。また，表 A·1 よりモデル間アダプターの全パラメータを拡張した Video Instruction Dataset のみで学習した Video-ChatGPT (w/o Init) と Act-ChatGPT (w/o Init) を比較すると，全体の正答率の差は僅かであった。このことから，Video-ChatGPT と提案モデルのモデル構造の違いは AutoEval-Video の評価に対しては大きな影響を及ぼさないといえる。したがって，この提案モデルにおける AutoEval-Video による評価の悪化は，提案モデルのモデル間アダプターの学習に使用されたデータの量及びその多様性が Video-ChatGPT のものと比較して，著しく劣ったことによるものであるといえる。

[*1] https://www.youtube.com/